DOCUMENT RESUME

ED 151 415 95 TM 006 944

AUTHOR	Donlon, Thomas F.; And Others
TITLE	Performance Consequences of Sex Bias in the Content of Major Achievement Batteries. Final Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C. Education and Work Group. Career Awareness Div.
REPORT NO	ETS-PR-77-11
PUB DATE	Jul 77
GRANT	NIE-G-74-0008
NOTE	149p.
EDRS PRICE	MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS	*Achievement Tests; Comparative Analysis; *Content Analysis; Elementary Secondary Education; Females; Item Analysis; Males; Scores; *Sex Differences; *Sex Discrimination; Sex Stereotypes; Standardized Tests; *Test Bias; Test Items; Test Results
IDENTIFIERS	California Achievement Tests; Iowa Tests of Basic Skills; Metropolitan Achievement Tests; Sequential Tests of Educational Progress

ABSTRACT

The content of selected major achievement tests is examined for bias in the frequency of references to the male and the female sex. The performance of males and females is contrasted on each item considered, and the relationships between item content and the performance differences is surveyed. The tests used in this study include the Metropolitan Achievement Test, the Iowa Tests of Basic Skills, the California Achievement Tests, and the Sequential Test of Educational Progress. Grades 2, 5, 8, and 10 are analyzed. Major findings include: (1) females perform better on items containing female references; (2) both males and females perform less well on items with a greater number of references; (3) consistent patterns occur regardless of whether the content variable studies is the total number of references or the total number of factors; (4) items with more than the average number of references do not differentiate ability levels within sex groups as sharply as do other items; and (5) both sexes are similar on such noncontent factors as omitting or rate-of-work. A Manual for Identifying Sex Bias Explicit in the Content of Test Items is appended. (Author/MV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151415

TM006 944

Grant No. NIE-G-74-0008

Final Report

PERFORMANCE CONSEQUENCES OF SEX BIAS IN THE CONTENT OF MAJOR ACHIEVEMENT BATTERIES

Thomas F. Donlon
Ruth B. Ekstrom
Marlaine Lockheed
with
Abigail Harris

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PR-77-11 July 1977

**EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY**

Performance Consequences of Sex Bias in The Content of Major Achievement Batteries

Thomas F. Donlon
Ruth B. Ekstrom
Marlaine Lockheed
with
Abigail Harris

Report Under

Grant Number NIE-G-74-0008

Prepared for: Career Awareness Division
National Institute of Education
U.S. Department of Health,
Education and Welfare

1977

Table of Contents

List of Tables v

List of Figures vii

Acknowledgements ix

Introduction 1

Methodological Considerations 5
- The Selection of Instruments 5
- Preparation of the Data Base 7
- Item Analysis Technique 9
- Methods of Content Analysis 13
- Comparison of the Content Analysis with that of Tittle . . . 16
- Quantifying Content in Related Material 17
- Coder Reliability 21

Results 23
- Item Characteristics Summary 28
- Correlational Analysis 36
- Sex Status and Performance 54
- Noncontent Factors 58
- Implicit Analysis 65
- Error Analysis 76

Summary and Discussion 81

References 85

Appendix A A-1

List of Tables

Page

Table 1 Broad Comparison of M/F Ratios for Tittle Study and Current Study 18

Table 2 Chi-Square Estimation of Reliability STEP Reading Form 2A 22

Table 3 Scott's π Reliabilities STEP Reading Form 2A 22

Table 4 Characteristics of the Item Analysis Samples 24

Table 5 Means and Standard Deviations for Selected Statistical and Content Characteristics 31

Table 6 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10 STEP Reading, Form 2A 37

Table 7 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10 STEP Language Arts, Form 2A 38

Table 8 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10 STEP Mathematics, Form 2A 39

Table 9 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10 STEP Social Studies, Level 2A 40

Table 10 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10 STEP Science, Level 2A 41

Table 11 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 8 ITBS, Vocabulary and Reading, Level 14 42

Table 12 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 8 ITBS, Language Skills, Level 14 43

Table 13 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 8 ITBS, Mathematics, Level 14 44

Table 14 Intercorrelations of Selected Content Factors and Item Characteristics, Grade 8 ITBS, Study Skills, Level 14 45

List of Tables (Cont'd)

		Page
Table 15	Intercorrelations of Selected Content Factors and Item Characteristics, Grade 5 ITBS, Vocabulary and Reading, Level 11	46
Table 16	Intercorrelations of Selected Content Factors and Item Characteristics, Grade 5 ITBS, Language Skills, Level 11	47
Table 17	Intercorrelations of Selected Content Factors and Item Characteristics, Grade 5 ITBS, Mathematics, Level 11	48
Table 18	Intercorrelations of Selected Content Factors and Item Characteristics, Grade 5 ITBS, Work Study Skills, Level 11	49
Table 19	Intercorrelations of Selected Content Factors and Item Characteristics, Grade 2 Metropolitan Achievement Tests; Reading, Primary II	50
Table 20	Percentage of Items at Each Grade Level in Three Status Categories	56
Table 21	Standardized Partial Correlation Coefficients for Variables in Equation for Each Grade	57
Table 22	Tests with "Omit" Response Greater than 1%	61
Table 23	"Not Reached" Comparisons Males Versus Females Grade 5	62
Table 24	"Not Reached" Comparisons Males Versus Females Grade 8	63
Table 25	"Not Reached" Comparisons Males Versus Females Grade 10	64
Table 26	Judges' Classification of STEP Reading Items for Implicit Sex Content	69
Table 27	Joint Distribution of Item Assignment by Judges and by Item Analysis	71
Table 28	Error Analysis for MAT Reading Primary II Percentage of Observed Wrong Answers for Each Possible Wrong Answer	72
Table 29	Error Analysis for STEP Reading 2A Percentage of Observed Wrong Answers for Each Possible Wrong Answer	74

List of Figures

Page

Figure 1 Sample Item Analysis for Females/Males California Achievement Test 12

ACKNOWLEDGMENTS

The writers wish to express their appreciation to the following associates at Educational Testing Service:

Kay Butler-Nalin and to Eleanor Clemson for their extensive assistance in preparing coding procedures and in carrying out the explicit content analyses.

Jerilee Grandy and Richard Harrison for the execution of the data analyses.

The study could not have been done without the cooperation of the major test publishers. Specific thanks go to the following persons who arranged for the transfer of data:

Dr. William Coffman, University of Iowa, on behalf of Houghton-Mifflin, publishers of the Iowa Tests of Basic Skills.

Dr. Harold Bligh, Harcourt, Brace, Jovanovich, Inc., on behalf of that firm as publishers of the Metropolitan Tests.

Dr. John Bianchini of the Berkeley Office of Educational Testing Service, publishers of the Sequential Tests of Educational Progress.

Dr. George Burket, CTB-McGraw-Hill, on behalf of that firm as publishers of the California Achievement Tests.

Special thanks go to Jessie Cryer and Irene Smith for the speed, care and patience with which they typed the materials.

Our appreciation to The National Institute of Education, Career Awareness Division and to Ms. Mary Lou Randour, Project Officer, for her interest in and support of this research.

Thomas F. Donlon
Ruth B. Ekstrom
Marlaine Lockheed
Abigail Harris

Performance Consequences of Sex Bias in The Content of Major Achievement Batteries

Introduction

A number of expectations concerning performance on educational and psychological tests are linked to sex stereotypes. Females, for example, are expected to do better on verbal tests, while males are anticipated to do better on mathematical tests. In this vein Maccoby (1973) has observed that "...during adolescence, females score on the average about a quarter of a standard deviation higher than males in verbal tasks," and Coffman (1961) remarked "If men and women make comparable scores on a particular test of verbal aptitude, one is likely to suspect some bias in the sampling of either men, or women, or both." Such statements reflect long-standing and well established expectations about differential behavior patterns on tests for the sexes.

Many of these expectations were substantially formed in the early years of the development of testing as a movement, and they tended to be confirmed in the data of that period. For example, in presenting the first reports of separate verbal and mathematical scores for the Scholastic Aptitude Test of the College Entrance Examination Board, Brigham (1932) found females surpassing males on the SAT-Verbal, by 31 scale points, while males outdid females on SAT-Mathematical by 26 points. The symmetry of these results, together with their consistence with established views of female and male interests, underscored the development of the expectations.

Increasingly, however, these expectations are being reevaluated in terms of the cultural influences which may determine them. Whereas, in the past, there was an implicit assumption that sex differences in test performance somehow reflect fundamental psychological differences between the sexes, it is increasingly evident that the cultural expectations themselves have operated in ways which tend to bring about the observed patterns. Thus, there is a self-fulfilling component to cultural expectations for sex differences. The increased awareness of this fact combines with a renewed concern for principles of social equality to focus attention on the implications of sex differences in test scores.

In a society which is formally dedicated to maximizing the development of individual potential, the proper determination of that potential must be given a high priority. Where test scores reflect and sustain cultural expectations which are in fact counter to the realities of human potential, this constitutes an obvious bias, one which may be in direct conflict with the ideals of the society. Given this larger social context, there is a vital need to understand the factors which contribute to the observed sex differences. The present study is a basic step toward securing this understanding.

A strong demonstration of the need to investigate sex differences in test scores is provided by data derived from administrations of the Scholastic Aptitude Test of the College Entrance Examination Board. The verbal SAT-V and its mathematical counterpart, SAT-M, are widely used to select individuals for post secondary education, often in highly competitive situations. To the extent that individual options for post-secondary education are influenced by the SAT, it is indeed a "major test." In the Scholastic Aptitude Test-Verbal of recent years there has been virtually no differences between the sexes in performance. Whereas 40 years ago there were the reported female superiorities of 31 scale score points, by the late 1960's females demonstrated only trivial superiorities of 1 to 3 scale score points (Angoff, 1971). Since the scale is so constituted that a single standard deviation is set at 100 points, the original Brigham data may be taken as roughly consistent with Maccoby's generalization, quoted above. But, as Angoff's data show, the generalization no longer applies, at least for this test. The reasons for this apparent shift in performance are not easily identified.

Test outcomes are the result of an interaction between test taker and test content. The content of tests, however, frequently reflects the very stereotypical expectations about sex and identity that prevail in the society. The pervasive effects of such stereotypes are, in fact, the fundamental basis for the expectations concerning sex differences in test performance. Thus, as in any stereotype, there is an obvious circularity, with test content itself serving to reinforce the acceptance of stereotypes which may affect test performance. Tittle (1973) and Lockheed (1973) have documented these stereotypes extensively.

Bias, of course, is present whether or not sex references in test content actually determine score outcomes. The simple presence of unequal numbers of references to each of the sexes is evidence of content bias. Certainly the qualitative distinctions among the referents, reflecting such sex stereotypes as "male laborer" and "female houseworker" constitute another form of bias. Lockheed (1973) has elaborated distinctions among these forms of referential bias, pointing to the association between sex and occupational role and sex and interests.

Studies of the impact of content bias on the relative performance of males and females have been reported by only a few investigators. Coffman (1961), Donlon (1973), and Strassberg-Rosenberg and Donlon (1975) studied the College Board Scholastic Aptitude Test. Hicks and Donlon (1976) studied the aptitude tests of the Graduate Record Examination. In each study, a number of demonstrations of the impact of content on the relative performances of females and males at the item level were reported, both for verbal and mathematical. Many of the verbal demonstrations were essentially in agreement with the stereotypical expectations. Thus, females did relatively better on Human Relations material, males on Science or Economics. Within mathematical material, however, some of the findings were unanticipated. There is a functional male advantage, but it is extensively modulated by content. Succinctly stated, females do relatively better on algebra than geometry. Demonstrations such as these at the item level do not in themselves establish the presence of test bias in the sense of an inequity in the distribution of scores or of decisions based on scores. But they clearly establish that the observed score differences between the sexes are a function in part of test content specifications, and that the magnitude of such content influences may be surprisingly great. Donlon (1973), for example, estimates that the observed sex differences of approximately 40 scale points on College Board Scholastic Aptitude Test - Mathematical, could be reduced to 20 or increased to 60 points by shifting the specifications with respect to the proportion of algebra and geometry items.

The present study examines the content of selected major achievement tests, analyzing it for bias in the frequency of references to the sexes.

Further, it contrasts the success of the sexes on each item in each of the tests considered. Finally, it surveys the relationships between these two factors, relating test content and the difference between the sexes.

Methodological Considerations

A number of methodological considerations underlie the work reported in this study. The major topics of this nature are reviewed in the following sections. Briefly, they are 1) the selection of instruments, 2) the preparation of the data base, 3) the conduct of the content analysis, 4) the relationship of the present content analyses to that of Tittle, 5) the approach to assessing content in material related to or associated with the items themselves, and 6) the assessment of coder reliability.

Methodological aspects of other parts of the study, such as the "implicit" analysis, or the comparison of wrong-answer distributions are covered in the sections which report the results of these activities.

The Selection of Instruments

The following criteria governed the selection of instruments for this study:

1. A focus on achievement batteries with a diverse set of content offerings. Content could not be restricted to any given subject such as mathematics.
2. An effort to select a total set of instruments which would span the entire grade range from 1-12. A number of researchers have reported age as a moderator of sex differences in interest

and attainment. The present study, while not designed to specifically determine these relationships, sought to compare the sexes at a number of points on the age and grade scale.

3. A preference for widely used and recognized batteries. In part, this study derives importance from the widespread use of the instruments it studies. Findings concerning obscure or seldom used instruments would be less compelling
4. A requirement that there be available national norms samples, for a year more recent than 1965, selected by procedures with a high level of statistical rigor.
5. The need to be able to secure item-level data on the individuals in the norms samples.
6. Preliminary evidence that test content in the items was sufficiently varied, in the sense of containing references to both males and females, that the effects of references to both sexes could be tested. The survey by Tittle (1973) was used to provide this preliminary evidence.

The application of these criteria led to the final choices:

Grades 9 - 12	Sequential Tests of Educational Progress, Series II, Level 2A, and California Achievement Test, Form A, Level 5 (Mathematical Problems Subtest only)
Grades 3 - 8	Iowa Tests of Basic Skills, Form 6
Grades 1 - 2	Metropolitan Achievement Tests (Reading only)

The Sequential Tests of Educational Progress (STEP) and the Iowa Tests of Basic Skills constituted major batteries with multitest, multilevel sampling. The mathematical STEP test, Mathematics Basic Concepts, however, does not contain much sex referential content. Accordingly, the Mathematics Problems Test of the California Achievement Test, Level 5, Form A was included as a supplement to the STEP test. Only this single subtest, Section 5 of the CAT battery, was studied.

The focus on the year 1965 or later as a criterion for the appropriateness of norms was a compromise between ideal samples and the realities of test publishing. Ideally, the more recent the sample, the better, for a great deal of social change with respect to sex role stereotypes has taken place since 1970 and is now continuing. In such circumstances the analyses of performances based on data gathered in 1968 or 1970 cannot be safely generalized to the present year. However, the opportunity to consider nationally representative samples on major batteries outweighed the constraints on generalization to the present.

A major decision on the availability of information was made in selecting the Iowa Tests of Basic Skills. There was considerable methodological interest in the study of responses to distractors or wrong answers in options. However, the item-level data for the ITBS battery is only available in terms of success or failure on the item. The decision to forego distractor information reflected the high standing of the ITBS on the other criteria. It was felt that adequate methodological demonstration of comparisons involving distractors could be made by examining the STEP tests and the Metropolitan Reading Test.

Preparation of the Data Base

The relevant test publishers--CTB-McGraw Hill, Educational Testing Service, Harcourt Brace Jovanovich, and Houghton-Mifflin--were approached with requests for data from the appropriate norms sample. In each case, complete cooperation was extended to this study. The request from the study was for a male and female sample, appropriately selected, with a minimum N of 1,000 in each sex. For the Iowa Tests of Basic Skills, this request was met precisely, because the very large Ns potentially available from their total norms sample dictated some kind of sampling. For the other publishers, however, it proved less expensive to transmit copies of the entire norms tapes at the appropriate level. This resulted in some variation in sample size from one major instrument to another, and from one subtest to another within the multitest major batteries. In each case, however, the required minimum of 1,000 cases was obtained. Further, additional information was in some cases supplied, such as both Primary levels of the Metropolitan Achievement Test.

The result was a potential for an empirical description of the relative difficulty for males and females in approximately 5,700 item-grade combinations, or approximately 11,400 basic item analyses. It was felt that this was potentially too large a component of project resources to devote to the statistical analysis, particularly since the assimilation of computer information from other systems into the ETS system, requiring somewhat special reorganization of tapes and layouts, was an unavoidable data processing expense. It was decided to limit the analysis to grades 2, 5, 8, and 10. This decision was based on a desire to consider the problem of sex bias across as broad a range of grades as possible. Further, it permitted a maximum use of the item content of the Iowa Tests. In the Iowa series adjacent grades take some overlapping item content. Grades 5 and 8, however, are sufficiently separated so that no item overlap occurred. This increases the total number of items evaluated.

Ultimately, then, the project performed approximately 3,000 analyses uniquely identifiable by item, grade level, and sex. These item analyses formed the basis for the correlations which constitute the major methodological approach of the study.

Item Analysis Technique

The item analysis was a standard program which provided the following data:

1. Frequency of selection for each response category.
2. Percentage of responses for each category.
3. Average score on the test, excluding the item itself, for each response category.
4. Standard deviation of scores on the test, excluding the item itself, for each response category.
5. Point biserial and biserial correlation coefficients

For the STEP and Metropolitan Tests the response categories were the optional responses, whose number depended on the specific characteristics of the item, and the Omit response or Not Reached response. For the Iowa Tests only four response categories were used: Right, Wrong, Omit or Not Reached.

The selection of a measure of item difficulty was based on practical and statistical considerations. It is a frequent practice in item analysis to substitute some sort of area transformation for the actual percentage, using the baseline values for the normal curve. One such transformation, for example, underlies the index of difficulty which is used in the standard item analysis at Educational Testing Service. This index, Δ, is the transformation of the observed percentage to the corresponding baseline value in a normal curve with mean 13 and standard deviation 4.

The principal advantage of such a change lies in the theoretical problems with the statistical operations of averaging or correlation. If the underlying distribution is a normal one, the percent passing will have a skewed sampling distribution. It was felt, however, that the sampling distribution for the item difficulties in the current study was likely to avoid extreme values, simply because the tests were carefully made to afford maximum discrimination. Throughout the range of percentages, from .10 to .90, the relationship between these values and a normal curve baseline value is not markedly non-linear, a point noted by Cureton (1963) and others. In the current study, therefore, this transformation was not applied.

Accordingly, the comparisons of item difficulty were in terms of percentages. The base for the percentage correct may be chosen from several options, all with some logical advantages and disadvantages. These are:

1. The total group of subjects.
2. The total group minus the number not reaching the item.
3. The total group minus both the number not reaching the item and the number omitting the item.

The distinction between "not reaching" and "omitting" is established in terms of whether any later item is responded to. Thus, if an item is not answered, but an item appearing later in the test is, then the item is considered to have been omitted. But if no later item is answered, then the item is considered to have been not reached. In effect, this assumes that the last item answered is the last item worked on, any later items being not reached. It implicitly assumes that blank or null responses prior to the last item are different from those which follow. They are generally held to be based upon a mental process in which the item is read but upon reflection not answered. The later items without responses are considered to have been unread, hence not reached.

None of the possible bases for the percentage passing is perfectly satisfactory, for the problem of dropout in item analysis is unavoidable. The calculation of percent pass based upon the number who reach the item is likely to provide an underestimate of the difficulty of the item for the total group, since those who fail to reach an item are typically less able subjects, with a lower than average probability of success. On the other hand, using the base of total subjects, regardless of progress through the test, opens the way to attributing differential item difficulty for the sexes to some content characteristic when in fact it might be due to differences in rate of work.

There are strategies for adjusting observed item difficulties for dropout. Thus, the standard ETS item analysis, which uses a normal curve baseline approach, adjusts the baseline measure by an increment or decrement appropriate for the shift in the average score of the group which actually responds to the item. Such adjustments, however, were considered too elaborate for the current study. Instead, the problem of Not Reached responses was approached for all items via a correlational analysis.

A sample item analysis for one item in the California Achievement Test is presented in Figure 1.

Figure 1

Sample Item Analysis for Females/Males - California Achievement Test

FEMALES

Item # 1 Difficulty = 0.8275 Point Biserial = 0.2504

Response Category	Total	1	2	3	4*	5	Omit	Out of Range	Not Reached
N	2357.0	95.0	21.0	120.0	1957.0	162.0	2.0	0.0	0.0
Per-Cent		4.03	0.89	5.09	83.03	6.87	0.08	0.0	0.0
Mean	8.101	5.179	3.905	7.133	8.417	7.253	8.000	0.0	0.0
Sigma	3.265	2.741	2.223	2.924	3.205	3.100	2.000	0.0	0.0

MALES

Item # 1 Difficulty = 0.8275 Point Biserial = 0.2504

Response Category	Total	1	2	3	4*	5	Omit	Out Range	Not Reached
N	2753.0	107.0	21.0	125.0	2278.0	221.0	1.0	0.0	0.0
Per-Cent		3.89	0.76	4.54	82.75	8.03	0.04	0.0	0.0
Mean	9.194	6.196	5.714	7.736	9.572	7.928	4.000	0.0	0.0
Sigma	3.306	3.257	2.881	3.015	3.199	3.117	0.0	0.0	0.0

Methods of Content Analysis

The approach to a content analysis sought to develop procedures which would be useful in analyzing any multiple-choice instrument. The task of the content analysis is to identify and to quantify differences in tests with respect to what they say about males and females, and the general concept of such a content analysis is almost intuitively clear: earlier studies by Tittle (1973) and Lockheed (1973) have established the general lines of the method. No earlier project, however, had provided a record of the procedures used for the content analysis which is sufficiently detailed to constitute a basis for replication of the study or its extension to new materials. This project sought to develop a record of its procedures sufficiently thorough for others to repeat them in some detail.

The complexity of language is a challenge to any system which seeks to describe it with quantitative analyses. Pronoun antecedents, for example, are often clearer to the writer than to the reader. Yet the identification of the proper antecedent is obviously necessary when the pronoun has to be counted as part of the person-referential content of the material. Proper names and mixed-sex plural groups also constitute special problems to a sex-oriented content analysis.

Added to the problems of language are problems of test and item structure. This study paid specific attention to item structure, both as a potential variable for additional analyses and as a characteristic which could assist in the rapid identification of differences in coding. But item structure varies from format to format. "Stem" and "options" may define a basic item concept, but there are a number of more complex situations than this, with response choices incorporated directly into material and designated by underlining, or with the stem and options only constituting a question in conjunction with some associated stimulus.

The content analysis procedures were as follows: For each item, an identification number was assigned designating the test, section, content of the section, and item number. Each item was also categorized as: (1) primarily verbal/text, (2) primarily pictorial, (3) primarily numerical/symbolic, or (4) mixed.

A test item is defined as having three components: (1) the stimulus, (2) the stem, and (3) the responses. Each of these components was analyzed separately.

The content analysis focused on four indicators: (1) the numbers of references to males and to females, (2) the numbers of male persons and female persons, (3) the various roles attributed to males and females in the content and (4) the relative status of male and female roles.

In order to determine the relative balance of males and females, the sex of actors must be identified. Four identification procedures were specified:

1. The noun is inherently sex-linked, e.g., mother father, sister, brother.
2. The noun is found to have a sex specific definition in the dictionary, e.g.,

   ballerina: 1. a principal female dancer in a ballet company, 2. any female ballet dancer.

3. The noun is a definite female or male name, e.g., Bill, Mary.

4. The noun has a male or female pronoun which refers to it, e.g., Pat went to her class.

The actor(s) in the item, either single individuals or a group of individuals, were then counted. In cases where sex could not be assigned, a neutral category was used.

In order to determine the relative balance of male to female references, the number of actors plus other words such as pronouns which refer to them in the item were counted; repetitions were also counted.

Words which showed vocations, avocations, or special functions of people (for example, doctor, mother) were coded as roles. Roles were not inferred from the descriptions of individual behavior. For example, the role of "househusband" is not inferred from the sentence, "He cleaned the house and fixed dinner." The identification of particular roles as female, male, or neutral was decided by the percentages of females and males found actively engaged in that role as documented by the Occupational Characteristics, 1970 Census of Population and other sources. When 80% or more of the individuals engaged in an occupation were one sex, the occupation was defined as a sex-typed role. Other occupations were classified as neutral roles. Historical consideration of roles was handled by general knowledge. If an item involved a role that was generally known as restricted to one sex (e.g., knights, the congressmen of 1800), it was coded as a sex-typed role.

In order to assess relative status, both males and females need to be present in an item. Assessing the status of roles in an item focused on the sex of the actor holding the most superior role. Thus, the status of males and females in an item could be equal, female superior, or male superior.

Comparison of the Content Analysis with that of Tittle

The content analysis carried out in the present study covered tests each of which had been considered by Tittle et al. in their earlier work. The technique of the present analysis was somewhat different, so that the results cannot be strictly compared. Further, the current analysis developed a statistical summary which focused on the item level, in contrast to the statistical description of total tests reported by Tittle. It is possible, however, to formulate the data from the two studies so as to reflect some of the consistency between them. Such a contrast can help to evaluate the extent to which the specific content analysis carried out in this project might distort content-difficulty relationships. The Tittle results reflected content only, but have been widely disseminated and were the first major and systematic application of content analysis to the problems of sex bias in tests. If the current analysis was greatly different from the Tittle results, the generalizability of the content-difficulty associations could be limited.

The major element of the Tittle approach was the ratio of male references to female references, considered in two classes: "all" and "regular." The former category included language which was masculine or feminine but in the generic sense as well as specific references to individuals or groups. The M/F ratios for this generic category may be contrasted with ratios developed by dividing "average male references" by "average female references" in the current study.

These M/F ratios are only approximates to the Tittle figures. Because the present study carried out an item analysis separately for each grade level, it did not incorporate all items in the statistical summary for the

ITBS Form Six results. However, there is substantial overlap in the material considered by the two approaches, and since the number being compared is a ratio, it is to a large extent independent of the size of sample of the items considered. Table 1 presents the comparison of the Tittle data and that of the present study. In general, the differences between analyses do not obscure a considerable comparability across tests. It is unlikely that the components of the content analysis used in this study have distorted the findings concerning content difficulty relationships so greatly that they cannot be considered compatible with the earlier results of Tittle.

Quantifying Content in Related Material

This study was concentrated on the item level, testing the relationships between content and difficulty which were obtained at that level. However, the quantification of content in order to assess its impact on item statistics is made difficult by the common practice in testing of basing multiple choice items on some sort of associated stimulus material. This associated material can be reading comprehension passages, graphs and tables, scientific diagrams, pictures, etc. It introduces a complication in the assessment of item content, for this content is associated with the total set of items but not part of any given item.

The nature of the problem is perhaps best seen by considering a hypothetical reading comprehension passage which makes six male references and three female references and which is the basis for four multiple choice questions. Clearly the content of the passage is relevant to the size of the differences between the sexes which are found with respect to success on these items. To show this relevance, the passage content

Table 1

Broad Comparison of M/F Ratios for
Tittle Study and Current Study*

Test or Subtest	Tittle Study	Present Study
STEP (Level 20)		
Reading	3.90	2.56
Language Arts	3.57	5.42
Social Studies	18.00	26.50
Mathematics	2.50	1.88
Science	14.00	-----
CAT Mathematics (Level 5, Form 6)	-----	-----
ITBS (Form 6)		
Vocabulary and Reading	7.31	6.12
Language Skills	2.11	1.78
Mathematics	1.76	2.12
Study Skills	1.82	1.57
Metropolitan Achievement Test Primary II	3.56	5.06

*The ratios for the present study consider most but not all of the items reflected in the ratio reported by Tittle.

must be allocated to the items in some way in order to carry out a sensible content-difficulty investigation.

One approach might be to consider the nature of the individual questions and to attempt to link paragraph content to individual items. For example, two of the male references and one of the female references might be attributed to a certain item depending on the judgment by a suitable expert that this apportioning reflects passage relevance to the problems posed by the given items. Such a model has certain advantages in terms of a sensible allocation of passage content. If the question is, in fact, answered on the basis of those sentences containing the specific content in question, then it seems plausible to consider that this specific content is the operative content in its impact upon item difficulty. A great deficiency of such a method, of course, would be the time required to make the expert judgments and the need to guard against overly subjective or unreliable decisions. Almost certainly a number of judges would be required. A possibly greater deficiency, however, lies in its concept of the mental process of the test taker. The method is correct in assuming that content will affect item difficulty if the test taker has an information gathering process that reacts to this content. But it seems more likely that the test taker's process is exploratory, and that it will range over the total passage. It seems unnatural, then, to allocate content only on the basis of its logical relevance to items. It seems more likely that the judgment of logical relevance made by a test specialist is a post facto decision, reached after reading the passage and the question, and is not an index of the consideration of content that the test taker considers in the process of responding.

Similarly, the likely process used by the test-taker makes it unsatisfactory to average passage content over the items. The questions are based on the passage. The passage as a whole is considered in responding, even if the process of evaluation is one of rejecting certain parts of it. The passage is not considered once, and then the item decisions reached. The passage is reconsidered in each decision. Passage content, then, must be considered to impact on each item.

This logic seems consistent with the handling of content in the various options of an item. Option content is summed to provide a contribution to total item content. It is not simply the content of the correct response that is evaluated, but the content of all of the options, as a set. This practice is clearly based on a model of test-taker process that sees the respondent considering all of the options before responding.

Given this model, the proper technique for quantifying item content is to sum the content of all the relevant components of the item situation: passage, item stem, item options. The impact of the content of stimulus material such as a reading passage is thus in a sense multiplicative. The passage content is seen as being considered by the test-taker each time an item is considered.

It is evident from this discussion that the impact of content on item difficulty is complexly affected by the model of the test taking process. An examination of such internal, process-oriented events, was beyond the scope of the present stuy, which analyses data gathered at prior administrations. The assumptions about process which have been made seem appropriate. But the consideration of tests at the item level is only beginning to receive appropriate attention. Studies such as those of Durost and Hodges (1974) seem likely to expand our knowledge of process and ways of mirroring it. Future studies of the test taking process may well reveal ways in which to improve the assumptions made here.

Coder Reliability

Complete double coding of all items was not feasible. Double coding was carried out for selected items for tests with significant content. These double codings were assessed using chi-square measures of distribution agreement presented in Table 2. In general, the values of chi-square are sufficiently small to support a high level of rater agreement. In an effort to place rater agreement on a scale more familiar to test users, the values of chi-square were supplemented by an index, π, proposed by Scott (1955) and defined as

$$\pi = \frac{P_o - P_\varepsilon}{1 - P_\varepsilon}$$

where P_o = the observed percentage of agreement and P_ε is the value expected by an hypothesis of chance overlap. It has a maximum value of 1.00 and a known sampling distribution giving it a greater resemblance to interrater correlation coefficients. As with chi-square, it establishes that raters agree with each other at a level of significantly greater than chance. Results for this index are presented in Table 3. The level of agreement among coders, as suggested by the data in Tables 2 and 3 was satisfactorily high.

Table 2

Chi-Square Estimation of Reliability
STEP Reading Form 2A

Part I (Items 1-30)

Unique Entities	x^2 = .028	(df = 3)	$p > .10$
References	x^2 = .025	(df = 3)	$p > .10$
Roles	x^2 = .713	(df = 3)	$p > .10$
Total Roles	x^2 = .0028	(df = 1)	$p > .10$

Part II (Items 1-30)

Unique Entities	x^2 = 1.754	(df = 4)	$p > .10$
References	x^2 = 1.212	(df = 4)	$p > .10$
Roles	x^2 = .144	(df = 5)	$p > .10$
Total Roles	x^2 = .312	(df = 2)	$p > .10$

Table 3

Scott's π Reliabilities
STEP Reading Form 2A

Part I (Items 1-30)

Unique Entities	π = .963
References	π = .970
Roles	π = .769
Total Roles	π = .870
Status	π = 1.000

Part II (Items 1-30)

Unique Entities	π = .853
References	π = .942
Roles	π = .912
Total Roles	π = .882
Status	π = .825

Results

Table 4 presents descriptive information concerning the tests and samples: the number of items, and comparative data, by sex, on the sample size, the means, the standard deviations and the internal consistency reliability estimates. In all, there were 36 subtests identified, each of which was item analyzed for males and for females. The minimum sample size used was 1,000 cases; the maximum was 3,384.

Sample size, of course, is an important factor in the determination of statistically meaningful differences. For $N_1 = N_2 = 1{,}000$, and for an average proportion in the two groups of 0.50, a difference in proportions of approximately .06 is significant at the .01 level. For the larger samples in the study, and for values of the average proportion other than 0.50, the minimum significant difference is even smaller. Thus, the sample sizes included for this study offer a sensitive statistical test of differences between the sexes.

The score means for the various subtests generally show equivalence between the sexes, with little evidence of superiority for one sex or the other. While many of these score differences achieve statistical significance, because of the large samples involved, their practical significance seems limited.

A number of the sex contrasts are in the directions suggested by stereotyping. Thus, significantly better performances are exhibited by females on both sections of the Metropolitan Achievement Tests: Reading and on the Iowa Test of Basic Skills Reading tests at Level 11 (Grade 5) and Level 14 (Grade 8). However, while females score consistently higher on the vocabulary tests for ITBS - Level 11, ITBS - Level 14 and STEP 2A, the differences are not statistically significant, and

Table 4

Characteristics of the Item Analysis Samples

Test Table		Number of Items	Sex	N	Mean	Standard Deviation	Reliability
Grade 10 Tests							
STEP 2A	Reading		F	1,129	22.1	5.0	.83
	Part I	30	M	1,070	22.1	5.3	.85
STEP 2A	Reading		F	1,129	12.4	6.0	.84
	Part II	30	M	1,070	12.2	5.8	.83
STEP 2A	Mechanics of		F	1,320	29.0	8.3	.89
	Writing, Part I	45	M	1,130	25.1	8.7	.89
STEP 2A	Mechanics of		F	1,320	16.0	13.7	.97
	Writing, Part II	45	M	1,130	15.0	12.5	.97
STEP 2A	English		F	1,329	22.6	6.4	.83
	Expression	40	M	1,168	21.5	6.6	.83
STEP 2A	English		F	1,329	10.3	5.7	.87
	Expression	25	M	1,168	9.5	5.2	.84
STEP 2A	Basic Math		F	1,303	24.4	7.2	.82
	Concepts	50	M	1,038	25.3	8.0	.85
STEP 2A	Social Studies		F	1,260	19.4	6.4	.84
	Part I	35	M	1,110	20.6	6.9	.86
STEP 2A	Social Studies		F	1,260	13.1	6.6	.85
	Part II	35	M	1,110	14.3	7.2	.87
STEP 2A	Science		F	1,248	23.2	6.7	.80
	Part I	45	M	1,237	25.4	7.6	.85
STEP 2A	Science		F	1,248	11.5	6.9	.90
	Part II	30	M	1,237	11.6	7.4	.91

Table 4 (cont'd)

Characteristics of the Item Analysis Samples

Test Table	Number of Items	Sex	N	Mean	Standard Deviation	Reliability
Grade 8 Tests						
ITBS Vocabulary	48	F	1,000	38.6	8.3	.87
		M	1,000	28.1	8.4	.86
ITBS Reading	80	F	1,000	49.5	13.4	.92
		M	1,000	47.7	14.5	.93
ITBS Spelling	48	F	1,000	27.7	9.4	.90
		M	1,000	22.6	9.6	.90
ITBS Capitalization	44	F	1,000	29.5	7.3	.86
		M	1,000	25.4	8.1	.87
ITBS Punctuation	44	F	1,000	26.6	7.9	.87
		M	1,000	22.2	7.9	.86
ITBS Usage	32	F	1,000	17.0	6.1	.84
		M	1,000	14.6	5.7	.80
ITBS Maps	42	F	1,000	24.1	7.3	.84
		M	1,000	25.4	8.0	.87
ITBS Graphs	28	F	1,000	17.0	5.1	.79
		M	1,000	17.2	5.5	.83
ITBS Reference Materials	59	F	1,000	37.5	9.6	.88
		M	1,000	33.7	11.0	.90
ITBS Math Concepts	48	F	1,000	28.2	8.9	.88
		M	1,000	28.0	9.2	.89
ITBS Math Problems	34	F	1,000	17.1	5.0	.74
		M	1,000	17.5	5.7	.80

Table 4 (cont'd)

Characteristics of the Item Analysis Samples

Test Table	Number of Items	Sex	N	Mean	Standard Deviation	Reliability
Grade 5 Tests						
ITBS Vocabulary	43	F	1,000	26.2	8.0	.87
		M	1,000	25.7	8.7	.89
ITBS Reading	74	F	1,000	47.0	11.1	.90
		M	1,000	45.0	13.2	.92
ITBS Spelling	43	F	1,000	26.7	8.0	.88
		M	1,000	23.0	8.6	.89
ITBS Capitalization	40	F	1,000	26.6	7.2	.86
		M	1,000	24.0	7.9	.88
ITBS Punctuation	40	F	1,000	28.0	7.1	.86
		M	1,000	24.8	8.0	.88
ITBS Usage	32	F	1,000	19.3	5.9	.81
		M	1,000	17.2	6.4	.84
ITBS Maps	36	F	1,000	22.6	5.8	.80
		M	1,000	23.2	6.5	.85
ITBS Graphs	26	F	1,000	17.3	4.8	.79
		M	1,000	17.3	5.5	.85
ITBS Reference Material	56	F	1,000	36.0	10.3	.90
		M	1,000	32.8	11.5	.92
ITBS Math Concepts	42	F	1,000	25.8	7.7	.87
		M	1,000	26.2	8.5	.90
ITBS Math Problems	29	F	1,000	17.0	5.2	.80
		M	1,000	16.4	5.9	.84

Table 4 (cont'd)

Characteristics of the Item Analysis Samples

Test Table	Number of Items	Sex	N	Mean	Standard Deviation	Reliability
Grade 2 Tests						
Metropolitan Achievement Tests, Primary II Sentences	13	F	3,265	9.8	3.2	.81
		M	3,384	8.9	3.4	.82
Metropolitan Achievement Tests, Primary II Stories	31	F	3,265	22.5	7.8	.93
		M	3,384	20.0	8.2	.92

this is also true for the Passages of STEP 2A Reading Part II. Thus, the female superiority on verbal material is centered more on reading comprehension than on vocabulary, and tends to be more prominent in the lower grades.

The more formal aspects of language usage: spelling, mechanics of writing, effectiveness of expression, etc., all favored females, with some of the largest intersex differences occurring in these areas.

For Mathematics, the pattern shows a statistically significant difference ($p < .01$) in favor of Grade 10 males on STEP Mathematics 2A; a statistically significant difference ($p < .05$) in favor of Grade 5 females on ITBS - Level 11 Mathematics Problems; but no other significant differences. These patterns of performance on mathematical tests are in accord with other studies which show that significant sex differences in favor of males do not appear until the high school years. On the test score level, then, the evidence of sex differences presented by this relatively large array of instruments held few surprises. The general effect of content on scores seemed to be about as described in earlier researches. The pattern seems to be (a) female superiority on verbal tasks throughout the elementary years, but no longer extending into the secondary level, (b) sex parity on arithmetic/mathematics in the elementary years, but a superiority for males during the secondary years.

Item Characteristics Summary

The present analysis was focused on the item level, rather than the test level. As a result of the detailed analysis, a number of descriptive characteristics were available for each item. These were logically divided into two classes. The first class included descriptions of the statistical characteristics of the items, such as the percentage of males or females succeeding, the difference in these percentages, etc. These descriptions were essentially the type of information which is referred to as item analysis in test construction. The second class included descriptions of the content characteristics of the items: the number of references to males, to females, to "neutrals" (people of unspecified sex), etc. The content analysis was, of course, an "item analysis" also; only convention reserves that label for a description of the statistical outcomes of administering the item to a sample.

The statistical item analysis included item-total correlations as well as the distributions of subjects across the possible responses. For these item-total correlations the "total" score was defined by the patterns of administration, and was based on those items administered together in a separately timed session. Thus, STEP Reading has, in all, 60 items, but the statistical analysis was performed separately for each 30 item section. This approach was a logically necessary step in considering the problem of dropout, or failure to complete, which is essentially defined only for a given timed session.

The level of generality of the test content used to define homogeneous item sets is therefore an important decision. The fundamental investigation centers on a measure of association between statistical and content characteristics, an association tested across items. Demonstrating an association for limited subsamples of items for reading passages, or for mathematics problems, was not a principal goal of the study. In a study involving many hundreds of items, the potential for such subdivision seemed endless.

Item material in tests is presented in a variety of manners and formats, all with a potential for interaction with sex differences and content factors. However, such very small sets of items do not constitute a large enough sample to do more than suggest relationships. The present study, therefore, sought to define meaningful groups of items which were sufficiently homogeneous in content to constitute a category, but did not seek to consider each of the many specialized item subgroups which could be defined. The consequence of this is, of course, that negative findings cannot mean that there are no statistical-content relationships but that at the level of analysis considered here, none were found.

Table 5 presents descriptions of the fourteen tests that were studied. For each set of items, the means and standard deviations of the item descriptors are presented. These descriptors are divided into the seven statistical characteristics and ten content factors.

The five statistical characteristics are:

1. Percent males passing, including Omits but excluding Not Reached. Thus, percent passing is defined as the Total Right divided by the Total Right plus the Total Wrong plus the Total Omits.
2. Percent females passing.
3. The difference in the percent passing, defined as Males-Females.
4. Male item-total biserial. This coefficient excludes the item itself from the total.
5. Female item-total biserial.

The ten content characteristics were:

1. Total male references
2. Total female references
3. Total neutral references
4. The difference between male and female references

TABLE 5

Means and Standard Deviations for Selected Statistical and Content Characteristics

Statistical Characteristics
STEP 2A

		% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Reading	M	.55	.56	-.01	.49	.49
	SD	.22	.22	.04	.18	.17
Language Arts	M	.49	.54	-.05	.56	.59
	SD	.17	.19	.05	.21	.23
Social Studies	M	.48	.45	.03	.48	.44
	SD	.16	.16	.04	.16	.16
Mathematics	M	.52	.49	.03	.42	.38
	SD	.19	.21	.06	.13	.14
Science	M	.49	.46	.04	.50	.44
	SD	.16	.17	.06	.19	.21

Content Characteristics
STEP 2A

		Total Male Ref.	Total Female Ref.	Total Neutral Ref.	Male-Female Ref.	Total Male Actors	Total Female Actors	Total Neutral Actors	Total Male Roles	Total Female Roles	Total Neutral Roles
Reading	M	17.48	6.83	14.03	10.64	3.15	.92	8.80	1.52	.97	3.78
	SD	21.89	19.35	13.75	24.93	3.19	1.91	11.24	1.84	2.22	5.11
Language Arts	M	.65	.12	.74	.53	.33	.08	.55	.25	.03	.33
	SD	1.39	.66	1.17	1.56	.70	.52	.88	.64	.16	.64
Social Studies	M	1.06	.04	8.30	1.01	.59	.07	5.37	1.11	.04	1.43
	SD	1.57	.27	9.45	1.57	1.00	.31	5.99	2.14	.27	2.39
Mathematics	M	.32	.17	.45	.15	.15	.08	.11	.09	.02	.11
	SD	1.17	.43	1.64	1.47	.59	.41	.36	.29	.12	.64
Science	M	.20	.00	.45	.20	.09	.00	.31	.08	.00	.13
	SD	.89	.00	.89	.89	.29	.00	.54	.27	.00	.41

TABLE 5 (cont'd)

Means and Standard Deviations for Selected Statistical and Content Characteristics

Statistical Characteristics
ITBS - 5

		% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Vocabulary & Reading	M	.60	.61	-.01	.49	.43
	SD	.16	.18	.05	.14	.13
Language Skills	M	.55	.63	-.07	.47	.46
	SD	.14	.15	.04	.14	.14
Mathematics	M	.58	.58	.00	.50	.44
	SD	.16	.17	.05	.12	.12
Study Skills	M	.62	.64	-.02	.48	.43
	SD	.12	.14	.06	.11	.12

Content Characteristics
ITBS - 5

		Total Male Ref.	Total Female Ref.	Total Neutral Ref.	Male-Female Ref.	Total Male Actors	Total Female Actors	Total Neutral Actors	Total Male Roles	Total Female Roles	Total Neutral Roles
Vocabulary & Reading	M	11.18	3.40	2.50	7.78	2.28	.68	1.74	1.23	.35	.78
	SD	17.12	10.76	3.03	21.44	3.20	1.76	2.04	1.73	.75	1.29
Language Skills	M	1.07	.99	1.32	.08	.68	.68	1.15	.32	.26	.19
	SD	1.66	1.93	1.63	2.48	1.02	1.53	1.45	.74	1.04	.97
Mathematics	M	.32	.35	.37	-.03	.14	.17	.13	.00	.08	.11
	SD	.92	1.34	1.58	1.69	.35	.70	.38	.00	.33	.36
Study Skills	M	1.03	.66	2.03	.36	.99	.64	2.89	.41	.11	.70
	SD	3.96	2.87	2.86	1.23	3.83	2.74	5.24	.82	.48	1.26

TABLE 5 (cont'd)

Means and Standard Deviations for Selected Statistical and Content Characteristics

Statistical Characteristics
ITBS - 8

		% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Vocabulary & Reading	M	.58	.60	-.02	.46	.44
	SD	.16	.17	.06	.15	.15
Language Skills	M	.48	.58	-.09	.44	.45
	SD	.15	.17	.05	.13	.14
Mathematics	M	.54	.53	.00	.42	.39
	SD	.15	.17	.05	.13	.17
Study Skills	M	.58	.59	-.02	.45	.41
	SD	.14	.15	.06	.11	.13

Content Characteristics
ITBS - 8

		Total Male Ref.	Total Female Ref.	Total Neutral Ref.	Male-Female Ref.	Total Male Actors	Total Female Actors	Total Neutral Actors	Total Male Roles	Total Female Roles	Total Neutral Roles
Vocabulary & Reading	M	13.19	.58	8.38	12.61	2.59	.20	3.27	1.17	.10	2.01
	SD	24.46	1.81	9.22	24.61	4.29	.62	3.83	1.51	.33	2.51
Language Skills	M	1.76	.60	1.40	1.15	.86	.34	1.22	.31	.24	.36
	SD	2.95	1.66	1.79	3.51	1.15	.84	1.35	.72	.73	.73
Mathematics	M	.57	.07	.20	.50	.24	.05	.07	.01	.01	.10
	SD	1.21	.38	.81	1.22	.51	.27	.31	.11	.11	.34
Study Skills	M	.36	.26	2.19	.11	.30	.17	1.42	.10	.02	.10
	SD	.87	.83	4.58	1.21	.74	.53	2.18	.30	.20	.30

TABLE 5 (cont'd)

Means and Standard Deviations for Selected Statistical and Content Characteristics

Statistical Characteristics
Metropolitan Achievement Test

		% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Reading	M	.63	.70	-.07	.64	.68
	SD	.09	.10	.03	.11	.15

Content Characteristics
Metropolitan Achievement Test

		Total Male Ref.	Total Female Ref.	Total Neutral Ref.	Male-Female Ref.	Total Male Actors	Total Female Actors	Total Neutral Actors	Total Male Roles	Total Female Roles	Total Neutral Roles
Reading	M	4.50	.89	2.90	3.61	1.86	.70	1.68	1.00	.52	.32
	SD	5.00	1.38	3.77	5.72	1.95	1.02	2.00	1.31	.90	.71

5. Total male actors
6. Total female actors
7. Total neutral actors
8. Total male roles
9. Total female roles
10. Total neutral roles

The averages for the item statistical characteristics reflect the same fundamental test qualities of difficulty level and reliability which are described in Table 4. Accordingly, no special discussion of them is necessary.

The averages for the content characteristics reflect some interesting patterns. These may be summarized as follows:

1. References to people are predominantly found in reading and vocabulary tests. There is a marked reduction in people references in mathematics or science.
2. Within the reading tests, references to males are about three times as numerous as references to females for Grade 10 and Grade 5 material, about five times as numerous for Grade 2, and about twenty times as numerous for Grade 8. Only once in fourteen tests did female references exceed male references, and this was in ITBS - Mathematics, Grade 5.
3. Neutral references in general show higher frequencies than females. Only in one test, ITBS - Vocabulary and Reading, Grade 5, did female references exceed neutral references. Neutral references exceeded male references on seven of the tests, but were exceeded by male references on seven others.
4. The numbers of different actors return these essential patterns. Female actors are much less frequently mentioned than males.

The totals of neutral actors exceed those of either males or females in ten of the tests, suggesting that the references-to-actors ratio for male references may be somewhat greater, since neutral references exceed male references in only seven of the tests.

5. The identification of roles is most difficult for any but the reading tests. This suggests that the descriptions of people in these other tests are very limited. In the four reading sets there are trends toward more roles per item in the tests for the upper grade levels. Male and female roles show a dip at Grade 8, but neutral roles increase systematically. Further, the ratio of neutral role increase is such that while they constitute only about 20% of roles at Grade 2 they constitute about 65% of roles at Grade 10. It seems appropriate to characterize these trends as more roles and more complex roles, for the higher grades.

Correlational Analysis

The selection of characteristics results in a 7 x 10 correlation matrix for each of the 14 achievement test sections. The results of these analyses are presented in Tables 6 through 19.

The most important statistical characteristic is the sex difference in percent passing. To what extent is this associated with content factors? Tables 6 through 19 provide 140 correlation coefficients of this variable with content. Of these, 28 or 20% are significant at the .05 level or greater under assumptions of bivariate normal distribution. This is not 20% of 140 independent relationships. In view of the

TABLE 6

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10

STEP Reading, Form 2A

N = 60 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	-.53**	-.48**	-.26*	-.45**	-.29*
Total Female References	-.27*	-.22*	-.27*	-.18	-.02
Total Neutral References	-.64**	-.61**	-.16	-.37**	-.25*
Male References Minus Female Ref.	-.25*	-.25*	-.02	-.25*	-.24*
Total Male Actors	-.59**	-.56**	-.18	-.36**	-.24*
Total Female Actors	-.32**	-.26*	-.32**	-.27*	-.08
Total Neutral Actors	-.51**	-.48**	-.17	-.20	-.07
Total Male Roles	-.47**	-.43**	-.25*	-.35**	-.20
Total Female Roles	-.32**	-.26*	-.32**	-.26*	-.07
Total Neutral Roles	-.49**	-.47**	-.14	-.29*	-.22

* = $p < .05$
** = $p < .01$

TABLE 7

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10

STEP Language Arts, Form 2A

N = 155 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	-.03	-.05	.07	.00	.05
Total Female References	-.04	-.02	-.06	.17*	.18*
Total Neutral References	.01	-.06	-.28**	-.07	-.06
Male References Minus Female Ref.	-.02	-.04	.08	-.08	-.03
Total Male Actors	-.03	-.03	.01	.11	.13
Total Female Actors	-.06	-.04	-.04	.14	.15*
Total Neutral Actors	-.11	-.17	.27**	-.01	.01
Total Male Roles	.03	.02	.04	.18*	.17*
Total Female Roles	-.01	-.04	.10	-.05	-.02
Total Neutral Roles	.01	-.02	.11	.03	.03

* = $p < .05$
** = $p < .01$

TABLE 8

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10

STEP Mathematics, Form 2A

N = 65 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	.13	.06	.22*	.16	.16
Total Female References	-.07	-.10	.11	.06	.12
Total Neutral References	.03	-.02	.16	-.06	-.01
Male Reference Minus Female Ref.	.15	.11	.11	.09	.06
Total Male Actors	.20	.13	.19	.20	.21*
Total Female Actors	-.08	-.11	.13	.07	.12
Total Neutral Actors	.01	-.04	.16	.01	.05
Total Male Roles	.12	.03	.27*	.15	.14
Total Female Roles	-.15	-.15	.07	-.07	-.10
Total Neutral Roles	.10	.10	-.01	-.05	-.05

* = $p < .05$
** = $p < .01$

TABLE 9

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10

STEP Social Studies, Level 2A

N = 70 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	.00	.04	-.15	-.05	-.02
Total Female References	-.05	-.04	-.01	-.11	-.14
Total Neutral References	-.05	-.07	.11	-.07	-.09
Male References Minus Female Ref.	.00	.05	-.15	-.03	.00
Total Male Actors	.01	.08	-.27*	-.14	-.13
Total Female Actors	.04	.06	-.09	-.12	-.11
Total Neutral Actors	.02	-.01	-.09	-.07	-.07
Total Male Roles	-.03	-.03	.04	-.09	-.12
Total Female Roles	-.05	-.04	-.02	-.11	-.14
Total Neutral Roles	-.14	-.15	.06	-.03	-.04

* = $p < .05$
** = $p < .01$

TABLE 10

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 10

STEP Science, Level 2A

N = 75 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	.20*	.15	.10	.13	.16
Total Female References	N/A	N/A	N/A	N/A	N/A
Total Neutral References	-.01	.00	-.03	.01	-.11
Male References Minus Female Ref.	.20*	.15	.10	.13	.16
Total Male Actors	.12	.17	-.15	.27**	.28**
Total Female Actors	N/A	N/A	N/A	N/A	N/A
Total Neutral Actors	-.10	-.01	-.24*	.09	.05
Total Male Roles	-.05	.04	-.26*	.20*	.22*
Total Female Roles	N/A	N/A	N/A	N/A	N/A
Total Neutral Roles	-.12	-.07	-.13	.00	.04

* = $p < .05$
** = $p < .01$

TABLE 11

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 8

ITBS, Vocabulary and Reading, Level 14

N = 128 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	.25**	.28**	.11	-.09	-.10
Total Female References	-.23**	-.20*	-.06	-.12	-.14
Total Neutral References	-.14	-.09	-.13	.08	.03
Male References Minus Female Ref.	.27**	.29**	-.11	-.08	-.09
Total Male Actors	.13	.20*	-.21**	.00	.01
Total Female Actors	-.23**	-.20*	-.06	-.12	-.14
Total Neutral Actors	-.06	-.02	-.13	.12	.07
Total Male Roles	.15*	.16*	-.05	.01	-.05
Total Female Roles	-.23**	-.20*	-.05	-.12	-.14
Total Neutral Roles	.17*	.22**	-.17*	.19*	.15*

* = $p < .05$
** = $p < .01$

TABLE 12

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 8

ITBS, Language Skills, Level 14

N = 168 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	.01	.01	.02	-.13	-.11
Total Female References	-.02	-.02	.01	-.09	-.03
Total Neutral References	.05	.07	-.11	.05	.06
Male References Minus Female Ref.	.02	.01	.01	-.07	-.08
Total Male Actors	.03	.01	.08	.21**	-.15
Total Female Actors	-.01	-.01	.00	-.08	-.03
Total Neutral Actors	.01	.02	-.02	-.03	.00
Total Male Roles	-.03	-.03	.02	-.08	-.06
Total Female Roles	.04	.03	.01	-.07	-.09
Total Neutral Roles	.02	-.01	.09	-.12	-.16*

* = $p < .05$
** = $p < .01$

TABLE 13

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 8

ITBS, Mathematics, Level 14

N = 82 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	-.02	-.12	-.01	-.04	-.06
Total Female References	.14	.09	.15	.10	.12
Total Neutral References	-.15	-.17	.10	-.12	-.16
Male References Minus Female Ref.	-.06	-.04	-.05	-.07	.10
Total Male Actors	-.13	-.11	.00	-.07	-.08
Total Female Actors	.16	.10	.16	.15	.16
Total Neutral Actors	-.18	-.18	.05	-.15	-.17
Total Male Roles	-.10	-.12	.08	.01	-.03
Total Female Roles	.14	.09	.14	.19*	.19*
Total Neutral Roles	-.22*	-.21	.03	-.14	-.18

* = $p < .05$
** = $p < .01$

TABLE 14

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 8

ITBS, Study Skills, Level 14

N = 129 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	.07	.08	-.04	-.01	-.01
Total Female References	.12	.09	.04	.06	.04
Total Neutral References	-.12	-.06	-.13	-.29**	-.22**
Male References Minus Female Ref.	-.03	-.01	-.06	-.05	-.04
Total Male Actors	.04	.06	-.07	-.04	-.03
Total Female Actors	.13	.11	.01	.07	.07
Total Neutral Actors	-.15*	-.03	-.24*	-.26**	-.17*
Total Male Roles	-.14	-.11	-.03	-.31**	-.28**
Total Female Roles	-.03	-.04	-.03	-.06	-.09
Total Neutral Roles	-.07	.00	-.15*	-.20*	-.17*

* = $p < .05$
** = $p < .01$

TABLE 15

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 5

ITBS, Vocabulary and Reading, Level 11

N = 117 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	-.20*	-.18*	.02	-.10	.11
Total Female References	.00	.07	-.23**	.07	.08
Total Neutral References	-.03	-.02	-.04	-.07	.05
Male References Minus Female Ref.	-.16*	-.18*	.10	-.11	-.12
Total Male Actors	-.16*	-.10	.14	-.06	-.03
Total Female Actors	.02	.09	-.25**	.04	.07
Total Neutral Actors	-.10	-.06	-.12	-.01	.00
Total Male Roles	-.06	-.05	-.01	-.09	-.05
Total Female Roles	-.02	.05	-.21	.06	.07
Total Neutral Roles	-.26**	-.19*	-.20*	.02	.00

* = $p < .05$
** = $p < .01$

TABLE 16

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 5

ITBS, Language Skills, Level 11

N = 155 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	.10	.11	-.07	-.01	.01
Total Female References	.04	.07	-.13*	.01	.03
Total Neutral References	.07	.03	-.13	-.22**	-.19**
Male References Minus Female Ref.	.03	.02	.06	-.01	-.02
Total Male Actors	.06	.07	-.05	-.03	-.01
Total Female Actors	.04	.06	-.10	-.04	.00
Total Neutral Actors	.06	.04	.07	-.20**	-.15*
Total Male Roles	.05	.03	.06	-.10	-.03
Total Female Roles	-.11	-.10	-.01	-.05	-.03
Total Neutral Roles	.10	.10	-.03	.04	.04

* = $p < .05$
** = $p < .01$

TABLE 17

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 5

ITBS, Mathematics, Level 11

N = 71 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	-.13	-.09	-.12	-.07	-.04
Total Female References	.21*	.29**	-.37**	.05	-.04
Total Neutral References	.08	.13	-.22*	.05	-.07
Male References Minus Female Ref.	-.24*	-.27**	.23*	-.08	.01
Total Male Actors	-.12	-.08	-.12	-.04	-.01
Total Female Actors	.21*	.27**	-.35**	.03	-.08
Total Neutral Actors	.14	.18	-.22*	.01	-.09
Total Male Roles	N/A	N/A	N/A	N/A	N/A
Total Female Roles	.19	.26*	-.34**	.07	-.04
Total Neutral Roles	-.06	-.03	-.07	-.14	-.13

* = $p < .05$
** = $p < .01$

TABLE 18

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 5

ITBS, Work Study Skills, Level 11

N = 118 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	.15	.12	.02	-.19*	-.13
Total Female References	.17*	.14	.01	-.19*	-.14
Total Neutral References	-.17*	.04	-.48**	.00	.05
Male References Minus Female Ref.	.08	.06	.02	-.18*	-.10
Total Male Actors	.15	.13	.00	-.18*	-.12
Total Female Actors	.17*	.15	-.01	-.18*	-.13
Total Neutral Actors	-.20*	.00	-.44**	.13	.13
Total Male Roles	-.05	.04	-.20*	-.20*	-.21**
Total Female Roles	.18*	.16*	-.01	-.15	-.15
Total Neutral Roles	-.11	.09	-.47**	.07	.01

* = $p < .05$
** = $p < .01$

TABLE 19

Intercorrelations of Selected Content Factors and Item Characteristics, Grade 2

Metropolitan Achievement Tests; Reading, Primary II

N = 44 Items	% Male Passing	% Female Passing	Sex Difference in Passing	Male Item Total Biserial	Female Item Total Biserial
Total Male References	-.05	.00	-.16	.22	.25
Total Female References	.04	.09	-.21	.08	.13
Total Neutral References	-.31*	-.32*	.15	-.21	-.18
Male References Minus Female Ref.	-.05	-.02	-.09	.17	.18
Total Male Actors	.07	.07	-.03	-.01	-.02
Total Female Actors	-.03	.01	-.17	.02	.05
Total Neutral Actors	-.03	.01	-.14	.11	.07
Total Male Roles	.20	.18	.03	-.06	-.06
Total Female Roles	.19	.23	-.22	.22	.24
Total Neutral Roles	-.01	.02	-.12	.31*	.30*

* = $p < .05$
** = $p < .01$

redundancy among content measures, there is increased likelihood of relationships for, say "total male actors," if "total male references" is significant. But even when attention is focused more narrowly on the total references only, approximately 20% of the coefficients achieve significance.*

In view of the frequency of significant, coefficients, it seems likely that there are relationships between content factors and sex differences in item success. The general level of these coefficients is quite low, with most being in the range of absolute values from .20 to .35. Five of the 28 significant correlations occur in the Grade 10 reading test. The others are spread more evenly over the remaining tests.

On the basis of these sex-difference correlations, there is some support for the hypothesis that total references, regardless of category (male, female, or neutral) are directly related to relatively greater female success. Of the 28 significant correlations, 23 are negative, indicating a sex difference more favorable to females for items with more references.

Each of the significant correlations with total female references is a negative value, indicating that the more female references there are in the item, the greater the relative success of females on that item. Since female references are much less frequently encountered than male references, the fact that their presence seems to enhance female performance, even to a very slight extent, is interesting. The results are too weak and irregular to posit a practical relationship. Such evidence as there is supports the view that females do somewhat better on items with greater numbers of female references; but note that the male and neutral references show such negative correlations in two of four cases, also.

---

*Significance levels used in this study are those deriving from normal bivariate relationships. A number of distributions encountered are sufficiently skewed to call this sampling theory into question.

Considering male and female rate of success separately, a number of quite strong relationships were found for STEP Reading 2A, Table 6. In general, these indicated that items with more references were harder both for males and females, and this was true regardless of the type of reference. The Grade 8 ITBS Reading Test also showed a number of significant relationships, but with somewhat a reverse tendency: the more the references to males or neutrals, the easier for males. Of the same relevant coefficients at Grades 2 and 5, all are negative. Again, the patterns are only suggested, but the weight of this evidence is that items with references are harder for both males and females. Females do relatively better, but both sexes found reference-rich items more difficult.

The impact on item-total biserial was studied also. Here the results seem so inconsistent from grade to grade and from test to test that summarization is difficult. Of the 140 male biserial relationships, 31 showed a significant level of association between item biserial and item content factor. Of these, 21 were negative, indicating that more references led to a lower biserial. This result could be artifactual, since more difficult items tend inherently to show lower biserials. (It is true that biserial _r_ is theoretically insensitive to item difficulty, and is certainly not constrained in its range in the way that phi or point biserial _r_ would be, but the effects of guessing serve to depress the mean of the correct response in a manner which lowers the coefficient.) There is some evidence, as stated earlier, that items with more references are harder.

While the same pattern is observed for females, the 22 correlations achieving significance are divided 13 negative, 9 positive. Clearly, the interpretation that items with more references were less highly correlated must be a guarded one.

The computation of item difficulty and of item-total biserial was based on groups which excluded the Not Reached subgroup. The presence of speed variance complicates sex-difference analyses in a number of ways. In most cases the item analysis shows that very few items incurred any dropout, with only the last third or so at all affected, and only the last tenth or so very strongly affected. A number of strong relationships between Not Reaching and content were demonstrated, both for males and for females. They tend to indicate that items with more references are the later items in a test, although there are some contrary coefficients. In general, the parallelism between males and females is such that no significant performance consequences would seem to arise from differential rates of work in the tests considered here. Where the significant correlations is demonstrated for males it is demonstrated for females also. It should be noted, however, that the correlations between the Not Reached index and the number of references tends to show that it is the later, more difficult items that have the most references. This finding must be kept in mind in examining the relationships between content and difficulty which were described earlier. It is true that items with more references tend to be harder, both for males and females, but this may not mean that they are harder because they have more references. As item writers create more difficult, complexly worded items, they may write longer ones, or items which they feel need "humanizing." One cannot analyze the data from the present study so as to resolve this question. Indeed, it would be difficult to design an experiment which sought to differentiate item difficulty and number of references so that the relationships could be clarified.

Sex Status and Performance

We have observed an imbalance of male references to female references within items. This imbalance alone constitutes content bias. It is, however, what this imbalance may suggest to the test taker that leads us to investigate the effects of content bias on test taker performance. That is, we are interested in determining whether the imbalance in the frequency of male references to female references is related to sex differences in item passing. Such a relationship would lend support to the idea that the test taker's interpretation of the appropriateness of the test content for himself or herself--based, in part, upon the actual observed frequencies of males versus females in the test items--will influence his or her motivation and hence achievement.

But raw or simple frequencies may be misleading when many items contain no references at all. A subsample of items containing any human reference was identified. This sample included all items in which there were references to women, men or both, or in which there were male, female or neuter entities, or in which there were any roles. This selection created a pool of 43 second grade items, 286 fifth grade items, 320 eighth grade items and 254 tenth grade items.

One critical issue which we hoped to examine was the relative status of the men and women in an item. Since so few items actually were coded as having status differences, a substitute for this variable was computed. Items in which male entities outnumbered female entities, male entities equalled female entities, and female entities outnumbered male entities were identified. Table 20 reports the percent of the items at each grade level having these characteristics. At all grade levels, items in which female entities outnumber male entities are the fewest, ranging from approximately 21% of the items at Grade 2 to approximately 5% of the items at Grade 10. Items in which male entities outnumber female entities, on the other hand, constitute the bulk of the items, ranging from approximately 54% of the items at Grade 2 to approximately 42% of the items at Grade 10. Only in this latter age range do the number of items in which males and females occur in equal numbers outnumber those in which males outnumber females (53.2% to 41.7%). At all other grade levels, items with more males outnumber those which treat the sexes equally.

In order to investigate the effect of this sex imbalance of content on test taker performance, a multiple regression analysis was undertaken. Other variables thought to have an effect on test taker performance were included along with two dummy variables representing the relative status of males and females in the item. Table 20 lists the variables and their measures. The variables fall into four large categories: achievement content area dummy variables, a dummy variable for a second source of test items, item characteristic variables, and the status variables.

From Table 21 it is evident that achievement content area and item characteristic variables are highly related to sex differences in item performance. The relevant comparison in the content area is the "language arts" content dummy.

Table 20

Percentage of Items at Each Grade Level in Three Status Categories

	Grade 2	Grade 5	Grade 8	Grade 10
More Female than Male Entities (dummy)	20.9	15.7	16.7	5.1
Female Entities = Male Entities (dummy)	25.6	42.0	35.9	53.2
More Male than Female Entities (dummy)	53.5	42.3	47.4	41.7
Total	100	100	100	100
N=	43	286	320	254

Table 21

Standardized Partial Correlation Coefficients for Variables in Equation for Each Grade

Percent of Males Passing Item Minus Percent of Females Passing Item

Item Content Predictors	Grade 2	Grade 5	Grade 8	Grade 10
Reading (dummy)	-----	.48***	.58***	.33***
Math (dummy)	-----	.13**	.32***	.44***
Social Studies (dummy)	-----	-----	-----	.57***
Science (dummy)	-----	-----	-----	.37***
Word Study Skills (dummy)	-----	.22***	.32***	-----
Test Identifier				
California Achievement Test (dummy)	-----	-----	-----	.07
Item Characteristics				
Other than Text Only (dummy)	.37**	.14**	.16***	-.07
Length of Item	.32*	-.10*	-.10**	-.07
Status Indicator				
More Female Entities than Male Entities (dummy)	.12	-.14**	.03	-.07*
Female Entities = Male Entities (dummy)	.27*	-.09*	-.10**	.06
R^2	.13	.20	.31	.41
N=	43	286	320	254

*p < .10

**p < .05

***p < .01

At all grade levels, however, significant sex effects are found. Because the direction of the effect is different between the second grade and all other grades, some caution in interpreting the effect must be taken. Since the number of items at the second grade level is so small, this effect may be an artifact.

At the fifth, eighth and tenth grade levels the effect is in the direction predicted. It must be held in mind that items in which more females pass the item than do males will have a negative number for the dependent variable. Thus, a higher percentage of females than males pass items containing more female than male entities or equal numbers of male and female entities. This is the case after the effects of item achievement content, format and item length have been partialled out.

Noncontent Factors

The primary focus in the study was concerned with the impact of content on the statistical properties of test items. These statistical properties, however, are influenced by the characteristics of the group taking the test, and it would be possible to err in attributing sex-linked performance differentials to content alone if the items were organized and presented in certain ways. That is, content-sex relationships might be an artifact of a more complex model in which content is correlated with some characteristic of test organization and this characteristic of the test is in turn related to differential performance by the sexes.

A good example of this possibility might be a situation in which the items with the greatest amount of sex-linked content are placed in the later stages of test as the last items. These items are the ones which are most influenced by variation among respondents in the rate of work. If the sexes differ in rate of work, fewer individuals of one sex will be working on these later items. The shift in number reaching may influence the statistical description of the items, and the stage would be set for a misattribution of statistical differences to content factors.

A similar difficulty is presented if there are sex differences in the tendency to omit. It is reasonable to assume that at any given level of ability the tendency to omit is a function of the difficulty of items: that harder items are more likely to be omitted than easier ones. But if the sexes differ markedly in a tendency to omit, then the sex with fewer overall responses will probably show less success. The other sex will show a greater percentage passing, for, while relatively large numbers of them do not know the answer, the incidence of correct responses due to chance alone may be appreciable. Thus, the more difficult the item, the greater the possibility of a sex difference due to differences in omitting. If the organization of the test is such that content-rich items tend in fact to be the more difficult items, the stage again is set for the misattribution of statistical differences to content differences.

Some aspects of these technical problems were reviewed in the discussion of the correlational analysis. It was felt, however, that potential sex differences in the tendency to omit or in the characteristic rate of work were in themselves variables of some interest. A number of psychological characteristics are attributed to the sexes in common stereotyping. Risk-taking, for example, is a characteristic more often attributed to males than to females. Since risk-taking might logically relate to the tendency to omit, or to the tendency to mark the last items in a test in a flurry of speed, an explicit description of both the Omit measure and the Not Reached measure was developed.

Table 22 shows all tests which had any item with a percent omitting for either sex greater than 1%. Any test not specifically cited in this table had no items with a percent omitting as great as this. For the tests which had some of these items, their frequency for males and females is given. Further, the subset of items which showed a tendency to omit greater than 2% is presented.

The general conclusion from this table and from the tests that had absolutely no such items is that the impact of omissions is trivial at any level, and that there is no demonstration of any sex difference in this tendency. It is true that the math tests show more frequent female omitting, but very few of these items reached the 2% level. There is certainly no support for a risk-taking differential between the sexes.

Tables 23, 24, and 25 provided descriptions of the Not Reached variable for males and females at two points in each test: the last item and the fifth from last item. While no standard index or convention has been established within the field of measurement by which to evaluate such data, the levels are such that most test workers would probably judge them to be unspeeded. The comparison of the sexes indicates that the level of difference between males and females is so small that no major impact on the items could be attributed to it. There is some evidence of correlations between content area and rate of work (more females reach later verbal items, for example, more males reach later mathematical items) but because the differences themselves are small, the patterns have little intrinsic interest.

Table 22

Tests with "Omit" Response Greater than 1%

Test	Total Items	Number of Items with Omit > 1% Males	Omit > 1% Females	Omit > 2% Males	Omit > 2% Females
California Achievement Test, Math Problems	15	3	7	0	2
STEP Reading Part II	30	1	9	0	1
STEP Mechanics of Writing Part II	45	0	1	0	0
STEP Math Basic Concepts	50	11	20	0	9
STEP Social Studies Part I	35	2	2	0	0
STEP Social Studies Part II	35	10	6	0	0
STEP Science Part I	45	6	7	1	3
STEP Science Part II	45	0	1	0	0
ITBS - Level 14 Reference Material	59	2	0	0	0
ITBS - Level 14 Math Concepts	48	5	4	1	0
ITBS - Level 14 Math Problems	34	5	10	0	1
ITBS - Level 11 Vocabulary	43	1	2	0	0
ITBS - Level 11 Maps	36	0	3	0	0
ITBS - Level 11 Math Concepts	42	3	6	0	0
ITBS - Level 11 Math Problems	29	1	2	0	0
MAT - Sentences	13	2	0	0	0
MAT - Stories	31	24	15	6	2

Table 23

"Not Reached" Comparisons
Males Versus Females

Grade 5

Content	% Not Reaching Last Item		% Not Reaching 5th from Last Item	
Iowa Tests of Basic Skills	Males	Females	Males	Females
Vocabulary	8.6	5.8	5.2	2.7
Reading	10.4	7.5	8.8	5.4
Spelling	13.7	8.1	8.6	4.3
Capitalization	8.4	4.3	4.7	2.1
Punctuation	2.6	1.1	1.4	0.0
Usage	2.3	0.0	1.1	0.0
Maps	19.2	23.9	10.4	12.0
Graphs	6.6	5.8	2.6	2.3
Reference Materials	17.6	13.8	14.9	10.8
Math Concepts	6.9	7.7	2.5	3.7
Math Problems	11.6	10.3	3.3	3.6

Table 24

"Not Reached" Comparisons
Males Versus Females

Grade 8

Content	% Not Reaching Last Item		% Not Reaching 5th from Last Item	
Iowa Tests of Basic Skills	Males	Females	Males	Females
Vocabulary	2.4	1.6	1.6	1.0
Reading	6.3	5.7	5.1	4.4
Spelling	8.5	4.9	5.5	3.2
Capitalization	4.2	0.0	2.3	0.0
Punctuation	1.6	0.0	0.0	0.0
Usage	0.0	0.0	0.0	0.0
Maps	11.8	14.2	7.1	6.5
Graphs	4.4	3.4	2.0	1.5
Reference Materials	7.3	3.7	5.0	2.3
Math Concepts	6.9	5.5	3.5	1.5
Math Problems	13.9	18.9	6.7	7.7

Table 25

"Not Reached" Comparisons Males Versus Females

Grade 10

Content	% Not Reaching Last Item		% Not Reaching 5th from Last Item	
Sequential Tests of Educational Progress	Males	Females	Males	Females
Reading Part I	2.3	2.4	0.0	1.0
Reading Part II	20.1	25.4	16.2	19.7
Mechanics of Writing Part I	1.4	0.0	0.0	0.0
Mechanics of Writing Part II	48.8	50.2	48.4	49.9
English Expression Part I	2.5	1.6	1.3	0.0
English Expression Part II	33.6	32.1	16.8	19.5
Math Basic Concepts	12.5	19.0	6.2	8.5
Social Studies Part I	3.6	4.2	1.3	1.8
Social Studies Part II	18.7	26.0	5.7	7.1
California Achievement Test				
Mathematics Problems	12.2	1.2	16.3	1.2

Implicit Analysis

The project centered on the explicit analysis of the tests, by which was meant that the measure of item content would be the frequency counts of explicit references. But it recognized the need for a consideration of what was called implicit analysis. Such an analysis could not be carried out for all of the thousands of items contained in the tests under review. Nor, indeed, would it be a meaningful process for judges in many cases. Much of the material was sex-linked only at very abstract levels of cultural knowledge. The ITBS material on Study Skills, for example, relating to Graphs, might be generally seen as more related to male interests and experience by application of a stereotypical logic concerning males and quantitative pursuits. But this would not be an implicit analysis of content in the sense suggested here. Accordingly, the implicit analysis was restricted to the Grade 10 sample and then to the reading material only. The reading material, richer in referential content, seemed a more appropriate focus for the activity.

The concept of an implicit analysis derives from the hypothesis that the content of tests may be sex-linked in ways other than through explicit references, such as through culturally determined opportunities to learn. Females tend to do better on aesthetic or artistic content, males on scientific or on mechanical knowledge. Thus, the proportion of tests which is devoted to these categories may be an important determiner of sex differentials. The studies by Coffman (1961), Donlon (1973), and Strassberg-Rosenberg and Donlon (1975) all demonstrate the potential utility of such implicit factors.

Eight persons were identified to carry out the implicit analysis. All were college graduates. There were five females and three males.

Recognizing the futility of generalizing from the sample of judges, a great effort was not made to reflect the characteristics of any population. The activity of the analysis was seen as a demonstration activity which tested the utility of the implicit approach less formally than might a statistical model.

The choice of highly educated judges, and implicitly, highly intelligent ones, seems sensible. What is being sought in this kind of analysis is the application of knowledge of a culture. This knowledge is held most strongly by those who by virtue of training and intellect have considered the impact of sexuality at length. But it is recognized that there are no perfect mirrors to culture. One must be in a culture to know many of its dimensions. Being in it, we surrender an essential objectivity.

The devising of appropriate instructions for the implicit analysis was difficult. While the judges were in general sophisticated about tests, the judgments which they were asked to make would draw upon the more tenuous characteristics of the language in tests. The instructions had to provide sufficient structure so that the judges' internal scales of the sex-referential dimensions of language were approximately in line with each other, but avoid prejudicing the group in the direction of perceiving a content imbalance of any kind. The following sample instructions were prepared:

> In connection with an NIE project on sex differences, we are conducting a comparison of the performance of males and females on a variety of achievement tests. Currently we are in the process of relating performance data to content characteristics. This involves an analysis of both explicit and implicit item content.

Content characteristics are explicitly determined by counting references to people: males, females, neutral. Thus, the item

> The boy and his father went to the zoo.

would be judged to have more explicit male content since it had two references to males and no references to female or neutral characters.

A different approach to content analysis relies more on the judgment of implicit sex differences. Thus, a vocabulary item

> The casserole was cooked too long.

has no explicit sex-referential content but might be judged to be more readily learned within female cultural experience. Judgments on implicit content should fall into one of the following categories:

1. masculine - content more readily learned within the male cultural experience
2. feminine - content more readily learned within the female cultural experience
3. neutral - content related to the same types of cultural experiences for both sexes
4. balanced - content favoring one sex are offset by other characteristics favoring the other sex

In some items, however, the analysis of implicit content will not be so clear-cut as in the preceding example. Sometimes the content of the item may be set in a masculine framework, while the skill required for successful completion of the item is more readily learned within the female cultural experience. Consider the following vocabulary item

> No mathematical theory was too abstruse for this man who had spent many years studying the subject.

Although the item deals with more masculine interests, the vocabulary word asked may be more familiar to women. In cases such as this, it is up to you to determine the predominant emphasis of the item.

Sensitivities to sex differences in cultural experiences vary with the age of the child, hence it is important to keep this dimension in mind when rating items. For the purposes of this analysis, you are being asked to consider this content with students in Grade Ten in mind.

> We would appreciate your help in rating the following items and reading passages. There are no right or wrong answers; it is merely a matter of personal judgment. As you develop questions and/or comments, we would appreciate them in writing to aid in the development of a brief "implicit analysis" manual.
>
> Thank you for your cooperation.

These instructions were responded to by some judges with appeals for greater definition. These appeals were received sympathetically with expressions of understanding for the difficulty of the task and the general problem of "What do you mean by the male cultural experience?" but the judges were encouraged to attempt the task aimed only with this level of definition.

The results of the analyses are presented in Table 26 which lists the judges' categorizations as Male, Female, Neutral, or Both. "Neutral," of course, is distinguished from "Both" in that it has, in the opinion of the judges, no sex-specific experiential background. "Both" implies a kind of separate but equal androgynous background.

These data may be evaluated from several viewpoints: Their internal consistency, their relationship to the explicit analysis, their relationship to the statistical properties of the items, and, finally, a comparison of the relationship of implicit and explicit analyses to the statistical outcome.

With respect to internal consistency it is evident that there is a generally high level of agreement among the judges. Some of this derives from their common tendency to designate Neutral responses. Such responses are designated in 269 of 480 instances. Particular attention must

Table 26

Judges' Classification of STEP Reading Items for Implicit Sex Content

	Part I					Part II			
	Male	Female	Neutral	Both		Male	Female	Neutral	Both
1	8	–	–	–	1	1	–	5	2
2	4	–	4	–	2	–	–	7	1
3	5	–	3	–	3	–	–	7	1
4	2	–	6	–	4	1	1	6	–
5	–	8	–	–	5	1	1	6	–
6	7	–	1	–	6	–	–	7	1
7	4	1	1	2	7	1	2	4	1
8	3	–	5	–	8	–	–	8	–
9	2	–	6	–	9	2	–	6	–
10	–	–	7	1	10	1	1	6	–
11	3	1	2	2	11	–	–	8	–
12	–	–	8	–	12	2	–	6	–
13	–	6	2	–	13	2	–	6	–
14	4	–	4	–	14	2	1	5	–
15	1	–	7	–	15	3	–	5	–
16	4	1	3	–	16	4	–	4	–
17	1	1	5	1	17	6	–	2	–
18	7	–	1	–	18	5	–	3	–
19	3	–	4	1	19	3	–	5	–
20	4	–	3	1	20	–	1	7	–
21	1	–	7	–	21	3	–	5	–
22	6	–	1	1	22	2	–	6	–
23	6	–	2	–	23	1	–	7	–
24	1	7	–	–	24	2	–	6	–
25	–	4	4	–	25	–	3	3	2
26	–	1	7	–	26	–	–	8	–
27	4	–	3	1	27	–	1	5	2
28	2	6	–	–	28	–	4	4	–
29	1	1	4	2	29	–	3	5	–
30	3	2	2	1	30	–	3	5	–

focus on cases of judgments which are diametrically opposite: one judge says the content implicitly favors males, another sees it as favoring females. For 13 items this was the case, but in all but two cases no more than one judge reached one of the two positions. Items 28 and 30 in Part I, however, drew M-F splits of 2-6 and 3-2 respectively. Each of these conflicts seemed to grow out of the juxtaposition of specific adjectives which were considered more frequently known by females, and nouns, which in the contexts of the stimulus sentences, were considered more frequently known by males.

The line of consistency between implicit analysis and explicit analysis is more difficult to summarize statistically. Each item is described by a vector of four numbers in the implicit analysis and by three in the "references" summation of the explicit analysis. To simplify, attention is focused on extreme cases. Eight items were judged masculine in content in the implicit analysis by five or more concurring judges. Six of those eight items were vocabulary items from among the thirty items in Part I. While four of these items were in the top third of all items in terms of male references, four had no male references whatsoever. Four items were judged female by five or more concurring judges in the implicit analysis. All were in Part I, the vocabulary section. Two of these had the strongest explicit content values; two had no reference at all. For these twelve items, accordingly, six had greater than average explicit values in the appropriate direction while six had no direct referential content at all. The judgments in the implicit analysis were apparently based on characteristics of the items not reflected in the explicit analysis.

The evaluation of the relationship between the implicit analysis and the difference in item difficulty between the sexes was made by converting judges' ratings into predictions. In order to do this 22 items were identified which had three or more judges indicating a judgment in favor of one sex. Item 30, Part I was not included in this analysis becuase of the evidence of conflict among the judges. Of these 22 items, 17 predicted male superiority, 5 predicted female superiority. The item analysis indicated that the 60 items were actually divided as 28 easier for males, 32 easier for females. Table 27 shows the resulting contingency table.

Table 27

Joint Distribution of Item Assignment
by Judges and by Item Analysis

Judges' Assignment

		F > M	M > F	M = F	
Item Analysis	F > M	4	7	21	32
	M > F	1	10	17	28
		5	17	38	60

$\chi^2 = 2.64$, $p > .10$

This contingency table indicates a weak association, not statistically significant, in support of judges' ability to predict the direction of item difficulty. This level of prediction is to be contrasted with the levels reported for the correlations of the explicit analysis. For the STEP Reading Tests these were generally in the low .20s and modestly significant. It indicates that the implicit judgments were probably not as predictive as the explicit ones. Judges were able to make decisions about the material, and to demonstrate a fair level of agreement among themselves, but there is limited evidence that the items they identify differ statistically from those they do not.

Table 28

Error Analysis for MAT Reading Primary II
Percentage of Observed Wrong Answers for Each Possible Wrong Answer

PART I					PART II				
Sentences					Stories				
		Answer Choices					Answer Choices		
Item	Sex	1	2	3	Item	Sex	1	2	3
1	M	-	47	53	1	M	-	73	27
	F	-	41	59		F	-	69	31
2	M	58	42	-	2	M	77	-	23
	F	57	43	-		F	81	-	19
3	M	-	63	38	3	M	52	-	48
	F	-	70	30		F	49	-	51
4	M	-	48	52	4	M	47	-	53
	F	-	38	62		F	45	-	55
5	M	54	-	46	5	M	32	68	-
	F	54	-	46		F	29	71	-
6	M	-	63	37	6	M	-	52	48
	F	-	64	36		F	-	52	48
7	M	53	47	-	7	M	44	56	-
	F	52	48	-		F	43	57	-
8	M	46	54	-	8	M	44	56	-
	F	49	51	-		F	42	58	-
9	M	44	-	56	9	M	-	53	47
	F	42	-	58		F	-	51	49
10	M	56	44	-	10	M	45	-	55
	F	64	36	-		F	38	-	62
11	M	67	-	33	11	M	-	46	54
	F	70	-	30		F	-	43	57
12	M	44	56	-	12	M	53	-	47
	F	43	57	-		F	46	-	54
13	M	40	60	-	13	M	-	57	43
	F	33	67	-		F	-	56	44

Table 28 (cont'd)

Error Analysis for MAT Reading Primary II

Percentage of Observed Wrong Answers for Each Possible Wrong Answer

PART II

Stories

Item	Sex	Answer Choices 1	2	3
14	M	67	33	-
	F	73	27	-
15	M	64	36	-
	F	75	25	-
16	M	50	-	50
	F	41	-	59
17	M	48	52	-
	F	58	42	-
18	M	-	72	28
	F	-	67	33
19	M	49	-	51
	F	42	-	58
20	M	66	-	34
	F	71	-	29
21	M	41	59	-
	F	44	56	-
22	M	69	-	31
	F	73	-	27
23	M	66	34	-
	F	73	27	-
24	M	-	49	51
	F	-	44	56
25	M	58	-	42
	F	57	-	43
26	M	-	43	57
	F	-	36	64
27	M	-	59	41
	F	-	60	40
28	M	61	39	-
	F	68	32	-
29	M	68	-	32
	F	69	-	31
30	M	-	63	37
	F	-	63	37
31	M	58	-	42
	F	64	-	36

Table 29

Error Analysis for STEP Reading 2A

Percentage of Observed Wrong Answers for Each Possible Wrong Answer*

PART I						PART II					
		Answer Choice						Answer Choice			
Item	Sex	1	2	3	4	Item	Sex	1	2	3	4
1	M	66	--	26	9	1	M	14	31	56	--
	F	67	--	27	6		F	12	34	55	--
2	M	--	28	29	42	2	M	25	--	53	21
	F	--	23	32	45		F	26	--	53	21
3	M	17	41	43	--	3	M	--	27	57	16
	F	29	30	42	--		F	--	30	50	20
4	M	30	42	--	28	4	M	--	51	24	26
	F	36	49	--	14		F	--	48	23	29
5	M	27	--	28	45	5	M	24	42	34	--
	F	13	--	32	55		F	28	44	28	--
6	M	--	42	35	22	6	M	30	20	50	--
	F	--	43	32	23		F	29	18	53	--
7	M	25	23	--	52	7	M	34	43	--	23
	F	29	24	--	46		F	33	41	--	26
8	M	--	29	41	30	8	M	--	18	33	49
	F	--	15	61	22		F	--	19	34	47
9	M	--	40	29	30	9	M	29	--	44	27
	F	--	41	21	38		F	34	--	35	31
10	M	--	52	36	13	10	M	--	55	33	11
	F	--	57	37	6		F	--	57	32	11
11	M	23	11	66	--	11	M	35	23	42	--
	F	24	5	72	--		F	39	20	41	--
12	M	13	74	--	13	12	M	27	32	40	--
	F	18	80	--	2		F	26	37	37	--
13	M	--	27	50	23	13	M	--	37	39	23
	F	--	19	52	29		F	--	35	39	26
14	M	21	60	--	19	14	M	16	50	--	35
	F	17	58	--	26		F	15	45	--	40
15	M	28	--	54	18	15	M	30	48	--	23
	F	34	--	54	12		F	34	44	--	22

Table 29 (cont'd)

Error Analysis for STEP Reading 2A

Percentage of Observed Wrong Answers for Each Possible Wrong Answer*

PART I						PART II					
		Answer Choice						Answer Choice			
Item	Sex	1	2	3	4	Item	Sex	1	2	3	4
16	M	32	28	40	--	16	M	19	31	49	--
	F	24	25	51	--		F	21	29	49	--
17	M	--	26	34	40	17	M	--	34	49	17
	F	--	19	28	52		F	--	39	44	17
18	M	33	--	23	43	18	M	30	--	42	28
	F	24	--	17	59		F	25	--	45	30
19	M	25	52	23	--	19	M	32	35	33	--
	F	26	54	21	--		F	31	38	31	--
20	M	--	31	47	22	20	M	31	--	38	31
	F	--	32	40	29		F	33	--	31	36
21	M	57	--	23	19	21	M	18	27	55	--
	F	53	--	21	25		F	15	21	63	--
22	M	--	22	63	15	22	M	21	--	30	50
	F	--	28	57	15		F	21	--	26	53
23	M	31	15	53	--	23	M	30	34	--	36
	F	29	13	58	--		F	23	34	--	43
24	M	66	--	26	8	24	M	--	29	37	34
	F	72	--	22	6		F	--	26	36	38
25	M	--	40	36	24	25	M	34	36	30	--
	F	--	37	38	25		F	40	28	32	--
26	M	29	42	29	--	26	M	35	34	--	30
	F	34	41	25	--		F	37	27	--	36
27	M	25	--	54	21	27	M	--	36	40	24
	F	23	--	57	20		F	--	35	40	25
28	M	28	40	32	--	28	M	24	50	--	26
	F	29	44	28	--		F	23	49	--	28
29	M	30	48	22	--	29	M	35	39	--	26
	F	31	49	19	--		F	36	37	--	27
30	M	11	--	80	9	30	M	24	41	35	--
	F	16	--	75	9		F	22	46	33	--

---

*Summed percentages may differ from 100 due to rounding.

In general, the judges were not comfortable with the processes. They sensed many of the ambiguities they faced. In fact, one judge explicitly reported an awareness of the conflict between feminine adjectives and masculine content. An example of this conflict might be "Their knowledge of algebraic transformations was precocious." Items and test content in general constitute a multi-faceted and complex communication. Judges could detect these various facets but their synthesis into an appropriate final categorization was not an easy one.

Error Analysis

The implicit analysis of the tests pinpointed the potential complexity of the item-person interactions. One demonstration of sex differences lies, of course, in the proportions selecting the correct answer to questions. But there are differences in the selection of the incorrect responses, also, and these can be a meaningful potential source of content relationships. The pioneer study of racial and ethnic differences by Davis and Eels (1950), for example, considered these differences. Tables 28 and 29 report such item comparisons.

Item 5 in STEP Reading 2A, Part I is of interest in regard to its error distribution. The item stimulus reads:

> "The other women thought Mrs. Watson was being extremely pretentious when she wore her new mink coat to the ball game."

The sexes do not differ greatly on this item in terms of level of success. The proportion of males succeeding is .921. The proportion of females is .927. However, as Table 29 reveals, there are differences in the relative attractiveness of the various misleads to this item. Fifty-five percent of the wrong-answering females select response 4, versus 45% of the erroneous males. Concomitantly, 27% of the males prefer response 1, versus 13% of females.

The content of these choices is of interest. Response 4, favored by females, is "uninformed;" response 1, favored by males, is "unfair." On the surface it would seem that there could be complex ways in which the different knowledges and frames of reference of the sexes are being brought to bear.

Item 4, response 4, is a wrong answer chosen twice as often by wrong-answering males (28%) as by wrong-answering females (14%). Again, the overall success of the sexes on the item is very similar: .851 for males, .858 for females.

Item 4 has the following stimulus:

> "In spite of the obvious yawns and vacant looks on the faces of his audience, the speaker launched into another interminable story."

Response 4 characterizes the speaker as "an expert." What characteristics of males or females could determine such results? Is the speaker seen as expert more frequently by males perhaps because he is male, or because he is boring? Both factors, of course, could operate. Such questions should be asked and answered in a context of significant differences. However, while there are several ways of evaluating such differences statistically, (Chi-square comparisons of wrong answer distributions, for example,) a difficulty with such analyses for the evaluation of content factors is the fact that the total Chi-square obscures the contribution of the individual cell. Further, the level of statistical significance is a function of the size of the sample. As noted in a number of earlier studies, it is often more important to responses as the main causal factor is implausible. The items as a whole provide a complex context which determines the interpretation of the options so that sex differences in response to words like "uninformed" or "unfair" are made operative only in this context.

The demonstration of a fairly high frequency of appreciable differences between the sexes in the selection of wrong answer options contributes to the understanding of item content factors as they relate to response. Sex differences should not be seen as arising from the stem only, or associated material, but from the item as a whole.

But the study of these sex differences cannot obscure the general equivalence of the sexes. Of the sixty reading items studied, there were only four cases where the most popular response for wrong-answering males failed to be the most popular response for wrong-answering females, only six cases where the least popular options were different. The fundamental picture is one of great similarity between the sexes in the response to written language regardless of content.

There are theoretical implications from this type of error analysis for an understanding of item difficulty; implications which must be examined in view of the number of studies which have focused on comparisons of item difficulties between the sexes. It might be conjectured, for example, that observed differences between the sexes in item difficulty could be a consequent of differences in the attractiveness of distracters. There is little formal understanding of the role of distracters in governing item difficulty. The practice of item writing, as reflected in published guidelines reflects the belief that widely remote distracters do not mislead and hence facilitate a respondent's correct response simply by reducing the domain of possible correct answers. If this reasoning is followed, differences in distracter attractiveness could theoretically account for differences in item difficulty. The fundamental question here considered, of course, is whether or not the choice of distracters is a sensitive element in the preparation of unbiased test material. Is there evidence that the characteristics of distracters govern or contribute to sex differences in the level of success? Because the study can only infer respondent thought process during item response, no definitive answer can be given. In many cases the overall difference in item success could be reduced to zero if the errors by one sex on a given distracter were reduced to the level of the other sex and the level of response to the correct answer incremented. But this model seems unjustified. It suggests that the power of a distracter lies in its capacity to induce in the respondent a mistaken conviction that this is the correct response. This might be considered a case in which differences in performance on distracters are related to differences between the sexes in misinformation.

More likely, however, it is not totally misinformation that governs response, nor does the respondent have strong convictions about the answer. Instead, the respondent has a state of uncertainty as to the correct answer, and the relative powers of the distracters lie in their attractiveness to these uncertain candidates.

One cannot rule out, therefore, the possibility that true sex equality in level of success on an item might be obscured by an apparent inequality which derived from differences among distracters in their attractiveness to uncertain respondents. It seems more likely, however, that while the content of distracters influences the choices of wrong answer differently for the sexes in some cases, it does not constitute the major factor.

Summary and Discussion

The major focus of this study was on the relation of sex differences in item difficulty to content factors in tests. The major findings would appear to be:

1. A significant tendency for females to do relatively better on items which contained female references. This was established by correlations with total female references across all items and by a regression analysis which focused only on items with male and female references. Only Grade 2 findings departed from this pattern.
2. A significant tendency for both males and females to do somewhat less well on items which had greater numbers of references.
3. Consistency in these patterns regardless of whether the content variable studied was the total number of references or the total number of actors. The total number of roles was a somewhat different, much less frequently applicable variable, whose pattern of correlation was somewhat more difficult to interpret.
4. Few references to people in tests other than reading or vocabulary.
5. A tendency for the number of references to be significantly correlated with the Not Reached variable, suggesting that items with more references occur later in the test. Since it

is a common test development practice to place harder items later in the test, this finding is congruent with the finding on total references and difficulty.

6. Items with more than average references do not differentiate ability levels within the groups, either males or females, quite as sharply as other items. This could be simply a reflection of their greater difficulty, for biserial indices tend to run lower for harder items.
7. The sexes are highly similar on such noncontent factors as omitting or rate-of-work. Such factors do not explain any of the observed content-difficulty relationships.
8. Judges ratings in an implicit analysis proved less effective at predicting sex difference than did the explicit measures of content. In general, the implicit analyses reflected quite different factors in item content.
9. Examination of wrong answer patterns revealed a small number of interesting examples of sex differences apparently related to the content of specific wrong answers and to sex differences in misinformation. Such findings could not diminish the overall impression of substantial agreement between the sexes, nor were the differences in wrong answers sufficiently large to explain the differences in item success.

As in any complex area of study, the investigation stimulates many new questions as it answers some old ones. It establishes significant but modest correlations between sex differences in success on items and the content of such items. While the level of these relationships, and the low overall frequency of explicit references to people would not make item content the prime source of sex differences, the finding is an interesting one. Since this study confirms the finding of Tittle that male references vastly outnumber female references in tests, the demonstration that female references modulate male-female difficulty differences has significant implications for test development practice.

Several limitations of this analysis should be discussed. First, there is the use of content analytic totals. These totals summed all references to individuals and to groups, obscuring potential differences in the impact of these two levels. Second, there is the redundancy in the concepts of references, actors and roles. This redundancy, while it can assist in the interpretation of patterns of coefficients, may also make it more difficult. Thirdly, the principal analysis considers only pass-fail differences in item response, ignoring the possible distinctions in wrong answer responding.

Some suggestions for future research would be:

1. To conduct studies of item-solution processes as these differ for the sexes. Little work of this type has been done. Connolly and Wantman (1964) offer the best example to date. Such a study would clarify and expand many elements of what is covered here.
2. To conduct studies of "quasi-parallel" forms of tests, forms parallel in every sense but with sex references varied in systematic ways. Such studies would provide clearer tests of the impact of content on difficulty.

3. To pursue the exploration of content subareas and sex differences, inverting the process used in the present study and seeking to establish the content characteristics of large-sex difference items. This approach characterized the earlier approach of Coffman. It may reveal test characteristics other than content related to sex differences in success.

It is, in one sense, reassuring to find the low level of statistical effects associated with content effects in this study. However, this finding must not be translated into a conclusion that attention to balanced sex references in tests is inconsequential. Tests are an important sampling of culturally significant behaviors. The conditions of this sampling should reflect the values of the culture. The patterns of sex differentiation are in transition, and the older practices of more frequent references to males, of stereotypical sex-role associations, and of sex and status linking should be abandoned. Bias in tests is not only demonstrated by statistical tests of the scores, but by the relative incidence of references to males and females.

References

Angoff, W. H. (Ed.). The College Board Admissions Testing Program. Princeton, N.J.: College Entrance Examination Board and Educational Testing Service, 1971.

Brigham, C. C. A Study of Error. New York: College Entrance Examination Board, 1932.

Coffman, W. E. Sex Differences in Response to Items in an Aptitude Test. Eighteenth Yearbook, National Council on Measurement in Education, 1961, pp. 117-124.

Connolly, J. A. & Wantman, M. An Exploration of Oral Reasoning Processes in Responding to Objective Test Items. Journal of Educational Measurement, June 1964, vol. 1, no. 1, pp. 59-64.

Cureton, E. E. Note on Vocabulary Test Construction. Educational and Psychological Measurement, 1963, vol. 23, pp. 461-466.

Donlon, T. F. Content Factors in Sex Differences on Test Questions. (Research Memorandum 73-28). Princeton, N.J.: Educational Testing Service, 1973.

Durost, W. N. & Hodges, R. B. A Study of Test-Taking Behavior for Two Independent Samples of Pupils at Grade Four with Special Emphasis on Guessing. Durham, N.H. Test Service and Advisement Center, 1974.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. Intelligence and Cultural Difference. Chicago, Illinois: University of Chicago Press, 1950.

Hicks, M. M., & Donlon, T. F. Sex Differences in Items Responses on the Graduate Record Examination. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, Calif., 1976.

Lockheed, M. Sex Bias in Educational Testing: A Sociologist's Perspective. Paper presented at the International Symposium on Educational Testing, The Hague, Netherlands, July 1973.

Maccoby, E. E., & Jacklin, C. N. The Psychology of Sex Differences. Stanford, Calif.: Stanford University Press, 1974.

Stafford, R. E. Sex Differences in Spatial Visualization as Evidence of Sex-Linked Inheritance. Perceptual and Motor Skills, 1961, vol. 13, p. 428.

Strassberg-Rosenberg, B., & Donlon, T. F. Content Influences on Sex Differences in Performance on Aptitude Tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D. C., 1975.

Tittle, C. K., McCarthy, K., & Steckler, J. F. Women and Educational Testing. Princeton, N.J.: Educational Testing Service, 1974.

APPENDIX A

A MANUAL FOR IDENTIFYING SEX BIAS EXPLICIT IN THE CONTENT OF TEST ITEMS

Kay Butler-Nalin
Marlaine E. Lockheed
Ruth B. Ekstrom
Thomas F. Donlon

with

Eleanor M. Clemson

Copyright © 1975, 1976, Educational Testing Service. All rights reserved.

A Manual for Identifying Sex Bias Explicit in the Content of Test Items*

Sex stereotyping in both the language and content of achievement tests has been noted by a number of researchers, including Coffman (1961), Donlon (1971), Lockheed (1973) and Tittle (1973). Sex stereotyping has been described explicitly as the absolute incidence of male and female references in the tests, as the relative status of males and females in the tests, as the incidence of sex-stereotyped roles in the tests, and implicitly as the sex-stereotyped content of the tests. In general, it has been found that the language of tests favors male-references over female references, depicts males in higher status roles than females, and describes males and females as stereotyped in both roles and characteristics.

What the effect of such stereotyping has been on the performance of test-takers is not known. Suggestions of the effect of test content on performance were found by both Coffman and Donlon, who report that test items discriminating between males and females statistically were found to have substantially different content.

The purpose of the present manual is to describe tests in such a way that the explicit stereotyping of language, roles and sex imbalance in content may be related to individual test performance. In order to relate sex stereotyping in tests to the actual test performance of males and females, it is necessary to select a unit of analysis most appropriate for this task. We have determined that the appropriate unit of analysis for this purpose is the test item itself. Although the manual was produced for a particular study, it can serve as a means of determining sex imbalance explicit in the content of any test.

A test item is defined as having three basic components: 1) the stimulus, which provides the background for the question, 2) the stem, which asks the question, and 3) the response(s), one of which is keyed as "correct." These item characteristics may be analyzed separately to determine their independent effect on the test taker's performance.

The explicit content analysis of the items in this manual focuses on three conceptually distinct but undoubtedly correlated indicators of sex imbalance. These are: 1) the relative balance of male references to female references, 2) the relative balance of males to females, and 3) the roles attributed to males and females.

By focusing on actors and roles in this content analysis, we are acknowledging our sociological bias. This aspect of the analysis entails asking whether the test taker's performance on the test is related to the extent to which the test taker perceives that the content of the test is appropriate for him or her. That is, we are asking whether role typification for males and females affects test performance. Role typification may be communicated by the balance of males and females, by the balance of male and female references, and by the actual portrayal of males and females in various roles.

---

*Support for the development of this manual came from a grant to Educational Testing Service from the National Institute of Education (NIE-G-74-0008) entitled "Performance Consequences of Sex Bias in the Content of Major Achievement Test Batteries."

Table of Contents

Page

Description of the Manual.................................. i

Coding Format.. 1

Coding Procedures.. 4
   General Information..................................... 4
   Category Types.. 5
   Counting People... 7
   Counting References..................................... 8
   Assigning Roles... 9
   Counting Roles.. 12
   Determining Status...................................... 13

Decision Rules... 14
   Identifying the Sex of Distinct Entities and References.... 14
   Counting Distinct Entities and References................. 15
   Designating Roles.. 18
   Miscellaneous Rules...................................... 20

Use of the Coding Sheet and Annotation Sheet............... 23

Table A - Words Which Are Inherently Sex Linked............ 26

Table B - Female, Male, and Neutral Roles.................. 27

Table C - List of Representative Collective Nouns.......... 32

Test Identification.. 33

Appendix

   Occupation List... 35
   Coding Layout... 39

Description of the Manual

The contents of this manual include four sections: (1) Coding Format, (2) Coding Procedures, (3) Decision Rules, and (4) Use of the Coding and Annotation Sheets.

Coding Format

The Coding Format section describes the codable components of each test item. As mentioned previously, these components are the stimulus, the stem, and the responses. The stimulus component of an item is the background information needed to answer the question. It may be stated directly in the question or assumed to be known by the test taker. The stem is the component of an item which demands a response. The responses are the possible answers to the question. Methods for identifying each component of the item are explained in this section.

Coding Procedures

The Coding Procedures section outlines the methods used to determine sex imbalance in item content. Sex imbalance may be defined as the relative balance of male and female actors, references to male and female actors, male and female roles, and the relative status of male and female roles.

In order to determine the relative balance of males and females, the sex of the actors must be identified. Four procedures are used:

1. The noun is inherently sex-linked, e.g., mother, father, sister, brother.

2. The noun is found to have a sex specific definition in the dictionary*, e.g.,

   ballerina 1. a principal female dancer in a ballet company, 2. any female ballet dancer

3. The noun is a definite female or male name, e.g., Bill, Mary.

4. The noun has a male or female pronoun which refers to it, e.g., Pat went to her class.

The actor(s) in the item, either single individuals or a group of individuals, are then counted. In cases where sex cannot be assigned, a neutral category is used.

In order to determine the relative balance of male to female references, the number of actors plus other words such as pronouns which refer to them in the item are counted.

---

*The dictionary selected for use for determining sex specific definitions was The Random House Dictionary of the English Language, 1967. A recently published dictionary was chosen on the belief that it would be less likely to contain sex-stereotyped definitions.

Words which show vocations, avocations, or special functions of people (for example, doctor, mother) are coded as roles. Roles are not inferred from the descriptions of individual behavior. For example, the role of "househusband" is not inferred from the sentence, "He cleaned the house and fixed dinner". The identification of particular roles as female, male, or neutral was decided by the percentage of females and males found actively engaged in that role as documented by the Occupational Characteristics, 1970 Census of Population and other sources. When 80% or more of the individuals engaged in an occupation were one sex, the occupation was defined as a sex-typed role. Other occupations were classified as neutral roles. Historical consideration of roles was handled by general knowledge. If an item involved a role that was generally known as restricted to one sex (e.g., knights, the congressmen of 1800), it was coded as a sex-typed role.

In order to assess status, both males and females need to be present in an item. Assessing the status of roles in an item focuses on the sex of the actor holding the most superior role. The status of males and females in an item can be equal, female superior, or male superior.

Decision Rules

The Decision Rules section provides directions for coding special situations. Particular problems have been found in identifying actors or groups, dealing with the structure of the English language, treating anthropomorphizing of animals, and handling ambiguous roles. In the manual these problems have been organized under the following headings which relate to the coding procedure: Rules for Identifying the Sex of Distinct Entities and References, Rules for Counting Distinct Entities and References, Rules for Designating Roles, and Miscellaneous Rules.

The Use of the Coding and Annotation Sheets

Special rules and procedures which are specific to the study "Performance Consequences of Sex Bias in Test Items" are covered in this section. It describes how the coding sheet is used, where on the coding sheet the content of the test questions are coded, and how irregularities and subjective coder decisions are documented.

CODING FORMAT

Test questions are coded by separating each item into stimulus (background information given or assumed known by the reader), stem (question which directly triggers a response), and response sections. For example:

stimulus ——→ The deep-sea diver went down twenty fathoms.

stem ——→ A fathom is equal to

response 1 ——→ A. 60 feet
response 2 ——→ B. 6 feet
response 3 ——→ C. 12 yards
response 4 ——→ D. 12 feet

Format I. Most test questions are easily analyzed into the three sections listed above and are frequently found in the following form:

Dale was five feet tall. (stimulus)

A foot is equal to (stem)

A. 2 inches (response 1)
B. 6 inches (response 2)
C. 9 inches (response 3)
D. 12 inches (response 4)

Format II. The stimulus, stem and response sections may not necessarily be listed in that order. For example:

The following sentence refers to which of the following meanings of the word "deserts"? (stem)

He received his just deserts when his job was given to someone else. (stimulus)

A. reward (response 1)
B. waste (response 2)
C. forsake (response 3)
D. punishment (response 4)

Format III. Frequently a reading passage, table, graph, map or picture is followed by a group of items referring back to it. In such cases, the passage, table, graph, etc., is coded as the stimulus and each new question referring to it is coded as a stem of that initial stimulus. For example:

O threats of Hell and Hopes of Paradise!
One thing at least is certain -- This life flies;
 One thing is certain and the rest is Lies;
The Flower that once has blown for ever dies. (stimulus)

The fourth line restates the second line (stem)

A. literally (response 1)
B. structurally (response 2)
C. onomatopeotically (response 3)
D. paradoxically (response 4)
E. figuratively (response 5)

In the fourth line "blown" means (stem)

A. blown up (response 1)
B. blown away (response 2)
C. bloomed (response 3)
D. died (response 4)
E. been planted (response 5)

Format IV. In some items the stimulus and response material may be the same and therefore double-coded. For example:

Which of the following are parts of a flower? (stem)

1. ovary
2. style } (stimulus)
3. pistil

A. 1 and 2 only (response 1) [Code as if reading text of 1 and 2]
B. 2 and 3 only (response 2) [Code as if reading text of 2 and 2]
C. 1 and 3 only (response 3) [Code as if reading text of 1 and 3]

Format V. Many items do not contain three sections or do not refer back to an initial reading passage, table, graph, etc. In such items, the stimulus is background information not given to the readers and is not coded. Code the question asked as the stem and code the responses as usual. For example:

Which of the following is not one of the uses of chlorophyll? (stem)

A. used to produce carbohydrates (response 1)
B. used as a disinfectant (response 2)
C. used as a dye (response 3)
D. used as a deodorant (response 4)

Format VI. Often in English expression and writing items, passages are presented which have the response choices incorporated into them. In such cases, code the entire passage as the stimulus, including the underlined sections designating response choices. Then code the underlined sections individually as response choices. (In such cases the stem is implicit and is not coded since no direct question is asked of the test taker.)

Example 1: John grabbed the ball from his friend than went
A B C
out to play. No error. (stimulus)
D

A = (response 1)
B = (response 2)
C = (response 3)
D = (response 4)

Example 2: In ancient Greek, large columns were a dominant
A B C
part of the architecture. No error. (stimulus)
D

A = (response 1)
B = (response 2)
C = (response 3)
D = (response 4)

Format VII. In spelling items, the test taker is often presented with alternative word spellings of a word to choose from. In such cases, simply code each choice as a response. For example:

A. scend (response 1)
B. send (response 2)
C. scent (response 3)
D. no error (response 4)

CODING PROCEDURES

Coding Procedures outline the methods used in coding the content of test items. It is divided into seven sections: general information, category types, counting distinct entities, counting references, assigning roles, counting roles, and determining status. Each of the thirty-nine coding categories is identified by name and mnemonic found on the coding sheet. (The mnemonic is in capital letters.) For further reference, please see the Sample Coding Sheet on page 24.

Section I: General Information

A. Test Identification ID

The number which identifies the test being coded is tallied in this category. The code for each test included in the present study is as follows:

1 = CAT (California Achievement Test)
2 = ITBS (Iowa Test of Basic Skills)
3 = MAT (Metropolitan Achievement Test)
4 = STEP (Sequential Tests of Educational Progress)

B. Section of Test SECTION

The number which identifies which part of the test being coded is tallied in this category. The code for each part is as follows:

1 = Part I
2 = Part II
3 = Part III
4 = Part IV
(etc.)

C. Section Content CONTENT

The kind of content in the section of the test being tallied in this category. The codes for the content of the section included in the present study are as follows:

1 = Reading Comprehension
2 = Language Arts (writing, grammar, spelling)
3 = Mathematics Problems
4 = Social Studies
5 = Science
6 = Work-Study Skills

D. Item Identification ITEM ID

The number of the test question or test item is coded in this category. The codes for the items or questions are as follows:

001 = test question 1
002 = test question 2

(etc.)

Section II: Category Types CATEG

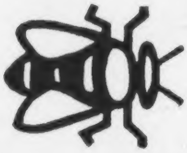
The coder makes the decision as to whether the item contains primarily (1) text or verbal, (2) pictorial, or (3) numeric or symbolic material and codes the type number according to the decision. Coder decision is based on the relative number of words, numbers, or pictures in an item, and on consideration of their function. The codes for the category types are as follows:

1 = primarily text/verbal
2 = primarily pictorial (includes maps, charts, cartoons with little or no textual material)
3 = primarily numeric/symbolic
4 = combination of two or more of the above

Example 1: Mary, Mary quite contrary; how does your garden grow:

Code: CATEG = 1

Example 2:



Code: CATEG = 2

Example 3: $5\ H_2O + 4\ CO_2 =$

Code: CATEG = 3

Example 4:

Insect	Life Span (days)
fly	100
mosquito	125
butterfly	175

Code: CATEG = 4

If the item does not contain one or more of the seven possible components (stimulus, stem, response 1, response 2, response 3, response 4, response 5) of the usual formats, a zero is coded in the CATEG category to show that the particular component is missing.

Example: Which is the color of the sun?

A. blue
B. brown
C. yellow
D. purple

Code Stimulus: CATEG = 0
Code Stem: CATEG = 1
Code each response : (A-D) CATEG = 1
Code fifth response: CATEG = 0

Section III: Counting People

A. Individuals: Female, Male, Neutral FEM, MAL, NUT

The number of distinct female, male, and neutral people in the item is tallied here. A noun which refers to people is coded as female or male: (1) if it is inherently sex linked (see Table A, page 26), (2) if it is found to have a sex specific definition in the dictionary (see Table A, page 26), (3) if it is a definite male or female name, (4) if it has a male or female pronoun which refers to it. When these conditions are not met, the noun is coded as neutral. Personal pronouns (such as I, we, you, she, he, etc.) are counted as distinct persons in the item when they are not references to a specific personal noun.

Example 1: The king and queen ruled well.

$$\text{Code: } \left.\begin{array}{l}\text{FEM = 1 (queen)}\\ \text{MAL = 1 (king)}\end{array}\right\} \text{ inherently sex-linked}$$

Example 2: She looked like her mother.

Code: FEM = 2 (she, mother) } inherently sex-linked

Example 3: The poetess was becoming famous.

Code: FEM = 1 (poetess) } sex specific definition

Example 4: The doctor grabbed her bag and rushed out the door.

Code: FEM = 1 (doctor) } female pronoun referent

Example 5: The nurse worked the night shift.

Code: NUT = 1 (nurse) } none of the 4 conditions above. [Note that because most nurses are female, this would be coded as a female role.]

B. Groups: Female, Male, Neutral (or mixed-sex) GPFEM, GPMAL, GPNUT

The number of distinct female, male or neutral (or mixed-sex) groups of people are tallied here. The same procedures apply to coding groups as female, male or neutral groups as to the coding of individuals in the previous categories. Both plural and collective nouns are coded as groups. A mixed-sex group is coded as GPNUT.

Example 1: The nuns went to vespers.

Code: GPFEM = 1 (nuns) } inherently sex-linked

Example 2: The 3 boy scout troops met after school.

Code: GPMAL = 3 (3 boy scout troops) } inherently sex-linked

Example 3: The team had a lot of spirit. } none of the 4 conditions

Code: GPNUT = 1 (team)

Example 4: Congratulations to the graduating students!

Code: GPNUT = 1 (students) } mixed-sex group

Section IV. Counting References

A. References to Individuals: Female, Male, Neutral FEMREF, MALREF, NUTREF

The number of references to female, male and neutral individuals in the item, including both nouns and pronouns, are tallied here.

Example 1: She loved her new car.

Code: FEMREF = 2 (she, her) } sex-linked pronouns

Example 2: John and Tom went to the show.

Code: MALREF = 2 (John, Tom) } male names

Example 3: Jack had his own car and Jill had hers.

Code: FEMREF = 2 (Jill, her) } female name and pronoun
MALREF = 2 (Jack, his) } male name and pronoun

Example 4: Pat was nominated to run for office.

Code: NUTREF = 1 (Pat) } name which could be either male or female

B. References to Groups: Female, Male, Neutral GPFEMREF, GPMALREF, GPNUTREF

The number of references to female, male and neutral (or mixed-sex) groups in the item are tallied here. When plural pronouns refer to groups of different sexes, the reference is coded as GPNUTREF. (GPNUTREF is a neutral reference or a mixed-sex reference).

Example 1: The crowd expanded until it poured into the street.

Code: GPNUTREF = 2 (crowd, it)

Example 2: The army was composed of brave men.

Code: GPMALREF = 2 (army, men)

Example 3: The ballet class had many students.

Code: GPNUTREF = 2 (class, students)

Example 4: Jack and Jill had their problems.

Code: GPNUTREF = 1 (their)

Section V: Assigning Roles

(For the present study, a role is defined as a vocation, an avocation, or a special human function.)

A. Individuals in Roles: Female Roles FEMROLF, FEMROLM, FEMROLN

The number of times that a female, male or neutral individual appears in a female role is tallied here. Decisions as to whether the role in the item is a female role can be aided by Table B (see page 27). When a female individual has a female role it is tallied in the FEMROLF category. If a male individual has a female role, it is tallied in the FEMROLM category. If a neutral individual has a female role, it is tallied in the FEMROLN category.

Example 1: Sally was hired as a nurse at a hospital nearby.

Code: FEMROLF = 1 (nurse)

Example 2: Being a prosperous hairdresser, he owned his own shop.

Code: FEMROLM = 1 (hairdresser)

Example 3: A flight attendant should not be too fat.

Code: FEMROLN = 1 (flight attendant)

B. Individuals in Roles: Male Roles MALROLF, MALROLM, MALROLN

The number of times that a female, male or neutral individual appears in a male role is tallied here. Decisions as to whether the role in the item is a male role can be aided by Table B (see page 27). When a female individual has a male role, it is tallied in the MALROLF category. When a male individual has a male role, it is tallied in the MALROLM category. When a neutral individual has a male role, it is tallied in the MALROLN category.

Example 1: As a judge, she worked long hours.

Code: MALROLF = 1 (judge)

Example 2: He became a minister.

Code: MALROLM = 1 (minister)

Example 3: Seeing a lawyer can be expensive.

Code: MALROLN = 1 (lawyer)

C. Individuals in Roles: Neutral Roles NUTROLF, NUTROLM, NUTROLN

The number of times that a female, male or neutral individual appears in a neutral role is tallied here. Decisions as to whether the role in the item is a neutral role can be aided by Table B (see page 27). If a decision cannot be made as to whether the role is male or female, it should be coded as neutral. When a female individual has a neutral role, it is tallied in the NUTROLF category. If a male individual has a neutral role, it is tallied in the NUTROLM category. If a neutral individual has a neutral role, it is tallied in the NUTROLN category.

Example 1: Brenda became an author at 16.

Code: NUTROLF = 1 (author)

Example 2: He enjoyed a long career as a psychology professor at a small college.

Code: NUTROLM = 1 (psychology professor)

Example 3: An actor experiences an up and down career.

Code: NUTROLN = 1 (actor)

D. Groups in Roles: Female Roles FEMROLFG, FEMROLMG, FEMROLNG

The number of times that a female, male or neutral (or mixed-sex) group appears in a female role is tallied here. Decisions as to whether the role in the item is a female role are aided by Table B (see page 27). Female, male, or neutral (or mixed-sex) groups in a female role are tallied in the same manner as individuals in female roles.

Example 1: The wives held interesting jobs.

Code: FEMROLFG = 1 (wives)

Example 2: The men were employed as elementary school teachers.

Code: MALROLMG = 1 (men)

FEMROLMG = 1 (elementary school teachers)

Example 3: The nurses went to their class.

Code: FEMROLNG = 1 (nurses)

E. Groups in Roles: Male Roles MALROLFG, MALROLMG, MALROLNG

The number of times that a female, male or neutral (or mixed-sex) group appears in a male role is tallied here. Decisions as to whether the role in the item is a male role are aided by Table B (see page 27). Female, male, or neutral (or mixed-sex) groups in a male role are tallied in the same manner as individuals in male roles.

Example 1: As generals, the women commanded respect throughout the ranks.

Code: MALROLFG = 1 (generals) FEMROLFG = 1 (women)

Example 2: The men went to classes to learn how to be bartenders.

Code: MALROLMG = 2 (men, bartenders)

Example 3: The taxi cab drivers were on strike.

Code: MALROLNG = 1 (taxi cab drivers)

F. Groups in Roles: Neutral Roles NUTROLFG, NUTROLMG, NUTROLNG

The number of times that a female, male or neutral (or mixed-sex) group appears in a neutral role is here. Decisions as to whether the role of a group is a neutral role are aided by Table B (see page 27). Female, male, or neutral (or mixed-sex groups) in a neutral role are tallied in the same manner as individuals in neutral roles.

Example 1: The female editors of McCall's did an article on women in publishing.

Code: NUTROLFG = 1 (female editors) FEMROLFG = 1 (women)

Example 2: The brothers were known as top paid actors.

Code: NUTROLMG = 1 (actors) MALROLMG = 1 (brothers)

Example 3: The garage workers and gas station attendants wanted higher wages.

Code: NUTROLNG = 2 (garage workers, gas station attendants)

Section VI: Counting Roles

A. Total Number of Female Roles TOTFEMROL

The total number of different female roles found in the item is coded in this category.

Example 1: My mother is a receptionist part-time and a nurse part-time.

Code: TOTFEMROL = 3 (mother, receptionist, nurse)

Example 2: Bill is a hairdresser and Ethel is a stewardess.

Code: TOTFEMROL = 2 (hairdresser, stewardess)

B. Total Number of Male Roles TOTMALROL

The total number of different male roles found in the item is coded in this category.

Example 1: The cab drivers and airplane pilots were on strike.

Code: TOTMALROL = 2 (cab drivers, airplane pilots)

Example 2: Studying to be an economist or a physicist is hard.

Code: TOTMALROL = 2 (economist, physicist)

C. Total Number of Roles TOTROL

The total number of the different roles, female, and male and neutral, are coded in this category. If only neutral roles are found in the item, their total is coded in this category.

Example 1: When my sister grows up, she wants to be a doctor, a lawyer, or a musician.

Code: TOTROL = 4 (sister, doctor, lawyer, musician)

Example 2: Once upon a time there was a princess who was lonely. She tried to find the jester in the palace to cheer her up but the jester was busy entertaining the knights and squires and princes so she stayed lonely.

Code: TOTROL = 5 (princess, knights, squires, princes, jester)

Example 3: The editors and reporters praised the sculptor's new work.

Code: TOTROL = 3 (editors, reporters, sculptor)

Section VII: Determining Status STATUS

The relative status of female and male roles in the item is coded in this category. The item has to have female roles and male roles present in order to be coded here. The role hierarchy is coded by focusing on the sex of the person holding the most superior role. If the item does not have both female and male present, it is coded as "not applicable." The coding of status is as follows:

1 = equal female and male status
2 = female status superior to male status
3 = male status superior to female status
4 = not applicable

Example 1: Bill and Mary are lawyers.

Code: STATUS = 1 (lawyers)

Example 2: The mother smiled proudly at her son.

Code: STATUS = 2 (mother > son)

Example 3: The daughter refused to be married although her father pleaded with her.

Code: STATUS = 3 (father > daughter)

Example 4: She can run faster than he can.

Code: STATUS = 4 (no roles)

Example 5: The doctor went home early.

Code: STATUS = 4 (only one role, doctor)

DECISION RULES

Special situations which arise within the previous thirty-nine categories are presented in this section, along with rules governing their coding. The Decision Rules for coding are divided into four sections: Identifying the Sex of Distinct Entities and References, Counting Distinct Entities and References, Designating Roles and Miscellaneous Rules.

Identifying the Sex of Distinct Entities and References

Rule 1. If sex is assigned to an individual or group at any point, carry that sex through the material.

Example 1: The dentist went to see his patient.

Code:
MAL = 1 (dentist)
NUT = 1 (patient)

MALREF = 2 (dentist, his)
NUTREF = 1 (patient)

MALROLM = 1 (dentist)
NUTROLN = 1 (patient)

Rule 2. When sex is assigned to an individual or group in an item, carry the sex through the stem and responses provided that the sense of personal identity is continued in the stem and responses.

Example 1:

Stimulus: Although the lawyer spoke forcefully, she did not sway the jury.

Stem: The lawyer can best be described as which of the following?

Responses:
A. The lawyer is a persuader.
B. The lawyer is a tactician.
C. The lawyer is a failure.

Code Stimulus:
FEM = 1 (lawyer)
GPNUT = 1 (jury)

FEMREF = 2 (lawyer, she)
GPNUTREF = 1 (jury)

MALROLF = 1 (lawyer)
NUTROLNG = 1 (jury)

Code Stem:
FEM = 1 (lawyer)
FEMREF = 1 (lawyer)
MALROLF = 1 (lawyer)

Code Each Response (A-C):
FEM = 1 (lawyer)
FEMREF = 2 (lawyer and persuader, or tactician or failure)

MALROLF = 1 (lawyer)
***ROLF = 1 (persuader, or tactician, or failure)

Rule 3. When forms of the verb "to be" are used, sex is carried through to the references, even when the sense of the sentence might be interpreted to indicate otherwise.

Example 1: She is a lawyer.

Code: FEM = 1 (she)
FEMREF = 2 (she, lawyer)

Example 2: She wants to be a lawyer.

Code: FEM = 1 (she)
FEMREF = 2 (she, lawyer)

Example 3: She never wants to become a lawyer.

Code: FEM = 1 (she)
FEMREF = 2 (she, lawyer)

Example 4: She pretended she was a lawyer.

Code: FEM = 1 (she)
FEMREF = 3 (she, she, lawyer)

Counting Distinct Entities and References

Rule 4a. If the number of individuals in a group is defined, or an actual number is given in the text, code the actual number of members in the Individuals categories; not in the Group categories.

Example 1: The twins had matching outfits.

Code: NUT = 2 (twins)

Example 2: There were eight orphans who needed homes.

Code: NUT = 8 (eight orphans)

Example 3: The sextet sang "Jingle Bells."

Code: NUT = 6 (sextet)

*** Consult Table B to determine whether *** = FEM, MAL or NUT

Rule 4b. When counting references in these cases, count the references as references to Groups, not Individuals.

Example 1: The twins had matching outfits.

Code: GPNUTREF = 1 (twins)

Example 2: There were eight orphans who needed homes.

Code: GPNUTREF = 1 (orphans)

Example 3: The sextet sang "Jingle Bells."

Code: GPNUTREF = 1 (sextet)

Rule 5. Code anyone/body, no one/body and someone/body as one individual (NUT = 1). Code everyone/body as one group (GPNUT = 1).

Rule 6. Code collective nouns* referring to people as being neutral groups (GPNUT).

Example 1: John was the only person in the audience who was listening.

Code: MAL = 1 (John)
GPNUT = 1 (audience)

Example 2: The union votes to strike.

Code: GPNUT = 1 (union)
GPNUTREF = 1 (union)

BUT COMPARE:

Example 3: Western civilization is decadent.

Code: 0

The difference between examples 2 and 3 is the focus on groups which can perform actions as a unit when counting distinct entities. The assumption made is that unions can act as a group while western civilization cannot.

Rule 7. Only subjects or direct objects determine the number of people or groups in an item. When linking verbs are used in a sentence, the predicate noun is not coded as another individual or a group.

Example 1: The wives were secretaries.

Code: GPFEM = 1 (wives)
GPFEMREF = 2 (wives, secretaries)

---

[*] See Table C for a list of collective nouns, page 32.

Example 2: Both men and women are elementary school teachers.

Code: GPFEM = 1 (women)
GPMAL = 1 (men)

GPFEMREF = 1 (women)
GPMALREF = 1 (men)
GPNUTREF = 1 (teachers)

Frequently the predicate noun is a role which describes the actor(s). The descriptor itself is not an additional distinct entity.

Rule 8. Code animals as individuals or groups, as references, and in roles (if applicable) when their appearance is personified or if they are anthropomorphized. Naming an animal implies anthropomorphism.

Example 1: The mother rabbit put all her baby bunnies to bed.

Code: FEM = 1 (mother rabbit)
GPNUT = 1 (baby bunnies)

FEMREF = 2 (mother rabbit, her)
GPNUTREF = 1 (baby bunnies)

FEMROLF = 1 (mother)
NUTROLNG = 1 (baby bunnies)

Example 2: The rooster pecked scornfully at the fence.

Code: MAL = 1 (rooster)
MALREF = 1 (rooster)

Example 3: Spot wagged his tail.

Code: MAL = 1 (Spot)
MALREF = 2 (Spot, his)

If animals are not personified or anthropomorphized, code only their pronoun references. Pronouns are coded because frequently animals are sex-typed.

Example 4: The dog buried his bone.

Code: MALREF = 1 (his)

Example 5: Lions kill in order to get their food.

Code: GPNUTREF = 1 (their)

Example 6: The deer ran into the forest.

Code: 0

Rule 9. Do not code inanimate objects or their references unless they are named or anthropomorphized.

Example 1: The family safe was well hidden.

Code: 0

Example 2: The bomb made the enemy weapons obsolete.

Code: 0

Example 3: Mother said, "Put G.I. Joe away and come to dinner."

Code:
FEM = 1 (mother)
MAL = 1 (G.I. Joe)

FEMREF = 1 (mother)
MALREF = 1 (G.I. Joe)

FEMROLF = 1 (mother)
MALROLM = 1 (G.I. Joe)

Example 4: Raggedy Ann is sleepy.

Code:
FEM = 1 (Raggedy Ann)
FEMREF = 1 (Raggedy Ann)

Designating Roles

Rule 10. Nouns which are inherently sex linked (see Table A, page 26) are coded as both individuals or groups and as roles in themselves. This rule does not apply to pronouns.

Example 1: The women went to work.

Code:
GPFEM = 1 (women)
GPFEMREF = 1 (women)
FEMROLFG = 1 (women)

Example 2: The brothers were mischievous.

Code:
GPMAL = 1 (brothers)
GPMALREF = 1 (brothers)
MALROLMG = 1 (brothers)

Example 3: She went to work.

Code:
FEM = 1 (she)
FEMREF = 1 (she)
FEMROLF = 0

Rule 11. When a role is used as a name, it is coded as a role.

Example 1: Dad went to work.

Code:
MAL = 1 (Dad)
MALREF = 1 (Dad)
MALROLM = 1 (Dad)

BUT COMPARE:

Example 2: Dan went to work.

Code: MAL = 1 (Dan)
MALREF = 1 (Dan)
MALROLM = 0

Rule 12. When a sentence contains only roles and not actual individuals or groups, code the roles in the total male roles or total female roles, and total roles categories Since such sentences do not contain actors, distinct individuals and groups are not coded.

Example 1: Studying to be an economist or a physicist is hard.

Code: TOTMALROL = 2 (economist, physicist)
TOTROL = 2 (economist, physicist)

Example 2: Careers combined with motherhood can be exciting.

Code: TOTFEMROL = 1 (motherhood)
TOTROL = 1 (motherhood)

Example 3:

Stem: Which word best describes a snoop?

Responses:
A. gossip
B. snob
C. saint
D. meddler
E. friend

Code stem: TOTROL = 1 (snoop)

Code each response (A-E): TOTROL = 1 (gossip or snob, or saint, or meddler, or friend)

Rule 13. When the subject of a sentence is a group which is subsequently described in role terms, count only the group or groups which is/are the subject of the sentence.

Example 1: The class was made up of nurses and doctors.

Code: GPNUT = 1 (class)

GPNUTREF = 3 (class, nurses, doctors)

FEMROLNG = 1 (nurses)
MALROLNG = 1 (doctors)

Example 2: Have you ever wondered what sort of people become volunteers?

Code: NUT = 1 (you)
GPNUT = 1 (people)

NUTREF = 1 (you)
GPNUTREF = 2 (people, volunteers)

NUTROLNG = 1 (volunteers)

Rule 14. Collective nouns do not carry roles. A role is defined as a set of expectations, beliefs, and behaviors that govern the action of an individual in a socially defined position. Thus, only individuals have roles.

Example 1: An actress was selected from the delegation.

Code: FEM = 1 (actress)
GPNUT = 1 (delegation)

FEMREF = 1 (actress)
GPNUTREF = 1 (delegation)

FEMROLF = 1 (actress)
NUTROLNG = 0

Example 2: The union voted to strike.

Code: GPNUT = 1 (union)

GPNUTREF = 1 (union)

NUTROLNG = 0

Miscellaneous Rules

Rule 15. When the stem and responses form a complete sentence but are incomplete in themselves, code the stem as if it would be coded if it were the beginning of a complete sentence and code each response as if it would be coded if it were at the end of a complete sentence.

Example 1:

Stimulus: Although the lawyer spoke forcefully, she did not sway the jury.

Stem: The lawyer was:

Responses: A. a persuader B. a tactician C. a failure

Code stimulus: FEM = 1 (laywer)
GPNUT = 1 (jury)

FEMREF = 2 (lawyer, she)
GPNUTREF = 1 (jury)

MALROLF = 1 (lawyer)
NUTROLNG = 1 (jury)

Code stem: FEM = 1 (lawyer)
FEMREF = 1 (lawyer)
MALROLF = 1 (lawyer)

Code each reponse (A-C): FEMREF = 1 (persuader, or tactician, or failure)
***ROLF = 1 (persuader, or tactician, or failure)

Example 2:

Stimulus: John went to work early so he could finish his assignment.

Stem: John is:

Responses:
A. a hard worker
B. class president
C. late for school

Code stimulus: MAL = 1 (John)
MALREF = 3 (John, he, his)

Code stem: MAL = 1 (John)
MALREF = 1 (John)

Code Response 1: MALREF = 1 (worker)
NUTROLM = 1 (worker)

Code Response 2: MALREF = 1 (class president)
NUTROLM = 1 (class president)

Code Response 3: 0

Rule 16. When man is not preceded by an article and is used in the generic, it is coded as GPNUT. In this usage, man does not have a role. Mankind is also coded as GPNUT, and does not carry a role. [This rule is an arbitrary decision made to simplify coding.] In an implicit analysis, words of this type would be considered as male.

Example 1: In one very important sense, man has suffered as a result of his technological sophistication.

Code: GPNUT = 1 (man)
MALREF = 1 (his)
GPNUTREF = 1 (man)
MALROLNG = 0

*** Consult Table B to determine whether *** = FEM, MAL, or NUT.

Example 2: Mankind has suffered because of technology.

Code: GPNUT = 1 (mankind)
GPNUTREF = 1 (mankind)
MALROLNG = 0

BUT COMPARE:

Example 3: A man becomes alienated when he immerses himself in the world of technology and automation.

Code: MAL = 1 (man)
MALREF = 3 (man, he, himself)
MALROLM = 1 (man)

Rule 17. When nouns or pronouns are used as possessives, code them as persons or groups, references, and roles if applicable.

Example 1: Her purse was stolen.

Code: FEM = 1 (her)
FEMREF = 1 (her)

Example 2: The boy's mother was angry.

Code: FEM = 1 (mother)
MAL = 1 (boy's)

FEMREF = 1 (mother)
MALREF = 1 (boy's)

FEMROLF = 1 (mother)
MALROLM = 1 (boy's)

Example 3: The family's safe was hidden.

Code: GPNUT = 1 (family's)
GPNUTREF = 1 (family's)

USE OF THE CODING SHEET AND ANNOTATION SHEET

This section defines special rules and procedures which are specific to the study, Performance Consequences of Sex Bias in Test Items. The first part of this section concerns procedures used on the coding sheet and the second part deals with special coding rules.

Coding Sheet

On page 24 is a sample of the coding sheet used in this study. The upper left-hand corner contains a box for coder initials. Beneath this box is a box for identification of test, subtest and page of the test being coded.

The coder begins coding with box one (1) under the General Information Section. The item information in the General Information section (ID, SECTION, CONTENT, ITEMID) is recorded only once for each item. The first line following the General Information section contains the codings in 35 content categories for the stimulus of an item. The second line contains the codings in 35 content categories for the stem of an item. The third line contains the codings in 35 content categories for Response 1 of an item. The fourth line contains the codings in 35 content categories for Response 2 of an item. The fifth line contains the codings in 35 content categories for Response 3 of an item. The sixth line contains the codings in 35 content categories for Response 4 of an item. The seventh line contains the codings in 35 content categories for Response 5 of an item. The eighth line is left blank and will later be filled in with data on performance on the item.

There is space to code two separate test items per coding sheet page. For future instruction as to the information coded into each individual box on the coding sheet, refer to pages 39-48.

Special Coding Rules

1. Each coder keeps an Annotation Sheet on which irregularities and subjective coder decisions are noted. The item in question is identified by the numbers which appear in the general information sections and notations or comments follow. See sample of an Annotation Sheet, page 25.

2. In the Distinct Entities and References sections there is room to enter double digits when counting. In the Roles section, there is only room to enter one digit per role type. If the number to be entered is greater than 9, code 9 in the box on the coding sheet, and note the actual number on the Annotation Sheet.

3. Note all items on the Annotation Sheet which have the word mankind or the generic man in them. See Rule 16, page 21.

4. Whenever the coder makes an arbitrary or subjective decision in coding an item, the decision should be recorded on the Annotation Sheet.

Coder initials:

PERFORMANCE CONSEQUENCES OF SEX BIAS IN TEST ITEMS

	test: subtest: pages:					DISTINCT ENTITIES						REFERENCES					
	General Info					People			Groups			People			Groups		
	ID	SECTION	CONTENT	ITEMID	CATEG	FEM	MAL	NUT	GPFEM	GPMAL	GPNUT	FEMREF	MALREF	NUTREF	GPFEMREF	GPMALREF	GPNUTREF
	1	2	3	4 5 6	7	8 9	10 1	2 3	4 5	6 7	8 9	20 1	2 3	4 5	6 7	8 9	30 1
stimulus																	
(stem)																	
(response 1)																	
(response 2)																	
(response 3)																	
(response 4)																	
(response 5)																	
stimulus																	
(stem)																	
(response 1)																	
(response 2)																	
(response 3)																	
(response 4)																	
(response 5)																	

	ROLES																				
	People.									Groups											
	FEMROLF	FEMROLM	FEMROLN	MALROLF	MALROLM	MALROLN	NUTROLF	NUTROLM	NUTROLN	FEMROLFG	FEMROLMG	FEMROLNG	MALROLFG	MALROLMG	MALROLNG	NUTROLFG	NUTROLMG	NUTROLNG	TOTFEMROL	TOTMALROL	TOTROL	STATUS
	2	3	4	5	6	7	8	9	40	1	2	3	4	5	6	7	8	9	50 1	2 3	4 5	6
stimulus																					
(stem)																					
(response 1)																					
(response 2)																					
(response 3)																					
(response 4)																					
(response 5)																					
stimulus																					
(stem)																					
(response 1)																					
(response 2)																					
(response 3)																					
(response 4)																					
(response 5)																					

SAMPLE ANNOTATION SHEET

Item #	Item Component	Notation
638001	Stimulus	The 9 coded in FEMROLF equals 13.
638005	Response 1	Man (generic) is used in this item.
638021	Stem	Coded braggart as NUTROL, role not found on Table B.

TABLE A

Words which are inherently sex linked that are automatically coded female or male plus words which are defined by the dictionary* as sex specific. (These words are also automatically coded as roles, see Rule 10, page 18.)

Female	Male
actress	boy
aunt	brother
ballerina	businessman
bride	cowboy
daughter	Dad
duchess	duke
girl	father
grandmother	fireman
lady	grandfather
Mom	husband
mother	king
niece	lord
nun	mailman
princess	male
queen	man
sister	nephew
stewardess	prince
wife	salesman
woman	son
	tycoon
	uncle

---

*The Dictionary used is The Random House Dictionary of the English Language (The Unabridged Edition), 1967.

TABLE B

Roles which have been designated as female roles, male roles or neutral roles. Roles were referenced from the occupation list (see page 35) . When 80% or more of the individuals engaged in an occupation were one sex, the occupation was defined as a sex-typed role. Other occupations were classified as neutral roles.

Female Roles

- airline stewardess/flight attendant
- aunt
- daughter
- duchess
- elementary teacher
- female
- girl
- grandmother
- hairdresser
- lady
- librarian
- Mom
- mother
- niece
- nurse
- office worker
- princess
- queen
- receptionist
- secretary
- sister
- wife
- woman

Male Roles

- airplane pilot
- ambassador
- banker
- bank officer
- bartender
- Board of Education member
- boy
- brother
- businessman
- clergyman, minister
- college president
- craftsman
- Dad
- dentist
- duke
- economist
- engineer
- farm laborer/farmer
- father
- fireman
- footman
- forester, conservationist/ lumberjack, firewatcher
- general
- geologist
- geophysicist
- governor
- grandfather
- Grandfather
- guard , watchman

Neutral Roles

- accountant
- actor
- anthropologist
- archeologist
- athlete
- author
- biologist
- clerk
- coach
- college professor
- curator
- editor and reporter
- elementary and secondary school administrator
- garage worker, gas station attendant
- inspector
- musician
- painter, sculptor
- psychologist
- psychology professor
- salesman
- sociologist
- sociology professor
- thief
- union member
- volunteer

TABLE B (Continued)

Female Roles	Male Roles	Neutral Roles
	historian	
	husband	
	janitor/building custodian	
	jeweler	
	judge	
	king	
	laborer	
	lawyer/attorney	
	lord	
	mailcarrier, postman	
	male	
	man	
	manager	
	mayor	
	nephew	
	pharmacist	
	philosopher	
	physicist	
	physician, doctor	
	plumber	
	policeman and detective	
	postmaster	
	prince	
	principal	
	railroad conductor	
	representative	
	robber	
	ruler	
	scientist	
	senator	
	son	

TABLE B (Continued)

Female Roles	Male Roles	Neutral Roles
	taxicab driver	
	uncle	
	veterinarian	

Roles which were not on the occupation list but appeared in the tests. The roles were generally classified as neutral roles except where otherwise indicated.

Female Roles	Male Roles	Neutral Roles
	alchemist	airline employee
	Captain	amateur
	city leader	apprentice
	city official	art collector
	Colonel	artist
	dictator	assistant
	footman	bore
	garbageman	boss
	jockey	cartoonist
	Lieutenant	citizen
	Little League Manager	cobbler
	marine	colonist
	mayoralty candidate	comedian
	member of the New York Stock Exchange	connoisseur
		council member
	merchant	coward
	political candidate	critic
	politician	demon
	Secretary of the Interior	employer
	Sergeant	enemy
	tycoon	experimenter
		expert
		explorer

TABLE B (Continued)

Female Roles	Male Roles	Neutral Roles
		factory worker
		fighter
		fisherman
		foe
		friend
		fugitive
		gangster
		hawker
		hiker
		hobo
		hunter
		immigrant
		institutional investor
		intellectual
		invader
		juror
		magician
		missionary
		monopolist
		motorist
		observer
		officer
		orphan
		outsider
		owner
		paint seller
		passenger
		peasant
		pewterer
		pioneer
		player
		poet
		president
		pupil
		ranger
		reader

TABLE B (Continued)

Female Roles	Male Roles	Neutral Roles
		referee
		reformer
		rogue
		scholar
		servant
		settler
		skater
		skipper
		slave
		socialist
		sorcerer
		specialist
		speaker
		student
		supervisor
		villain
		wholesale furniture dealer
		worker

TABLE C

List of Representative Collective Nouns

union
government
army
delegation
convention
organization
labor
management
corporation
community
culture
civilization
society
colony
empire
population
audience
band
class
committee
company
crowd
gang
group
majority
mankind
team

Test Identification of Subtests Coded for "Performance Consequences of Sex Bias in Test Items"

Numbers assigned follow the order of the General Information Section. Test ID is the first number. Section of test is the second number. Content of the section is the third number. [Item number is the fourth through sixth number and not given in this list].

STEP (Series II) Tests

411 = Reading, Form 2A, Part I
421 = Reading, Form 2A, Part II

412 = Mechanics of Writing, Form 2A, Part I
422 = Mechanics of Writing, Form 2A, Part II
432 = English Expression, Form 2A, Part I
442 = English Expression, Form 2A, Part II

413 = Mathematics Basic Concepts, Form 2A, Part I
Mathematics Computation, Form 2A is not being coded.

414 = Social Studies, Form 2A, Part I
424 = Social Studies, Form 2A, Part II

415 = Science, Form 2A, Part I
425 = Science, Form 2A, Part II

MAT (Form F) Tests

331 = Primary I, Test 3, Reading, Part A: Sentences
341 = Primary I, Test 3, Reading, Part B: Stories
371 = Primary II, Test 3, Reading, Part A: Sentences
381 = Primary II, Test 3, Reading, Part B: Stories

ITBS (Form 6) Tests

211 = Test V, Vocabulary (levels 9-14)
221 = Test R, Reading Comprehension (levels 9-14)

212 = Test L, Language Skills, L-1: Spelling (levels 9-14)
222 = Test L, Language Skills, L-2: Capitalization (levels 9-14)
232 = Test L, Language Skills, L-3: Punctuation (levels 9-14)
242 = Test L, Language Skills, L-4: Usage (levels 9-14)

216 = Test W, Work-Study Skills, W-1: Map Reading (levels 9-14)
226 = Test W, Work-Study Skills, W-2: Reading Graphs and Tables (levels 9-14)
236 = Test W, Work-Study Skills, W-3: Knowledge and Use of Reference Material (levels 9-14)

213 = Test M, Mathematics Skills, M-1: Mathematics Concepts (levels 9-14)
223 = Test M, Mathematics Skills, M-2: Mathematics Problem Solving (levels 9-14)

CAT (Level 5, Form A) Tests, 1970 Edition

113 = Section 5, Mathematics Problems

APPENDIX

Occupation List

The identification of particular roles as female, male, or neutral was decided by the percentage of females and males found actively engaged in that role as documented by several different sources. The following is a list of occupations found in the tests being coded, the percentage of females actively involved in that occupation, and the source of the percentage. The primary sources and their notations appear below. Other sources are listed along with the occupations.

* Occupation Characteristics, 1970 Census of the Population, "Summary of Social and Economic Characteristics of the experienced civilian labor force by detailed occupation and sex: 1970"

** World Almanac, 1975

++ Occupational Outlook Handbook (1974-75) U. S. Department of Labor, Bureau of Labor Statistics, 1975 Bulletin.

Occupation	Percentage of Females	Source
actors	39	*
airplane pilots/aviators	1	*
airline stewardess/flight attendants	95	*
ambassadors	1	**
anthropologists	53.8	Monthly Labor Review, Nov. 1975, Vol. 98, #1 Proportion of Doctorates Earned by Women, By Major Field of Study and Subfield, United States, 1969-1972
archaeologists	26.83	Monthly Labor Review, Nov. 1975, Vol. 98, #1 Proportion of Doctorates Earned by Women, By Major Field of Study and Subfield, United States, 1969-1972
athletes	27	*
authors	29	*
bank officers/bankers	6	*
bartenders	20	*
biologists	34	*

Occupation List (Continued)

Occupations	Percentage of Females	Source
Board of Education Members	12	Women on School Boards, National School Board Association Report #1974-1.
clergymen/ministers	2	*
clerks	25.2	Monthly Labor Review, Nov. 1975, Vol. 98, #1 Table 6, Employment by occupation, total and women workers, 1962-1974
coaches	25	*
college presidents	5	Survey done in fall '75 of accredited colleges & universities by the Office of Women in Higher Education of the American Council on Education
craftsmen/craft workers	4.5	Monthly Labor Review, Nov. 1975, Vol. 98, #11 Table 1, Occupational distribution of employed women, 1962-1974
curators/archivists	30	*
dentists	3	*
economists	11	*
editors and reporters	40	*
elementary and secondary teachers	69	*
elementary teachers	83	*
engineers	1	*
farm laborers/farmers	16	*
firemen	1	*
foresters, conservationists/ lumberjacks, firewatchers	3	*
garage workers, gas station attendants	28	*
generals	1	**
geologists	3	*
geophysicists	2.91	Monthly Labor Review, Nov. 1975, Vol. 98, #1 Proportion of Doctorates Earned by Women, By Major Field of Study and Subfield, United States, 1969-1972

Occupation List (Continued)

Occupations	Percentage of Females	Source
governors	1	**
guards	4	*
hairdressers	31	*
historians	8	*
inspectors/checkers	50.6	Monthly Labor Review, Nov. 1975, Vol. 98, #1 Table 3, Occupational participation rates: women as a percent of the total employed workers, 1962-1974
janitors	25	++
jewelers	11	*
judges	4	*
lawyers/attorneys	4	*
mailmen/postmen	8	*
managers/managerial	18.5	Monthly Labor Review, Nov. 1975, Vol. 98, #11 Table 1, Occupational distribution of employed women, 1962-1974
mayors	1	U.S. Conference of Mayors, 1620 Eye Street, N.W., Washington, D.C.
musicians	35	*
nurses	97	*
office workers/clerical	68.8	Monthly Labor Review, Nov. 1975, Vol. 98, #11 Table 1, Occupational distribution of employed women, 1962-74
painters, sculptors	35	*
pharmacists	11	*
philosophers	6	Survey done in 1974 by the American Philosophical Association
physicians/doctors	9	*
physicists	3	*
plumbers	1	*

Occupation List (Continued)

Occupations	Percentage of Females	Source
policemen and detectives	3	*
postmasters	2	*
professors (college)/faculty	28	*
psychologists	38	*
psychology professors	29	*
principals	elementary 19.6 jr. high 2.9 sr. high 1.4 total: 13.5	"Women in Administrative Positions in Public Education" position paper from Recruitment Leadership Training Institute, Temple University. 1974
railroad conductors	1	*
receptionists	97	*
representatives (House)	3	**
robbers/burglars	7	**
rulers/heads of countries	5	**
salesmen	41	*
scientists	13	*
secretaries	97	*
senators (federal)	0	**
sociologists	45	*
taxi cab drivers	6	*
thieves	7	**
total laborers (except farm)	8	*
union members	24	**
veterinarians	5	*
volunteers	57	Americans Volunteer, Washington, D.C.: Action, Feb. 1975, p.24.

CODING LAYOUT

GENERAL ITEM INFORMATION

Box 1	ID
Box 2	SECTION
Box 3	CONTENT
Boxes 4-6	ITEMID

STIMULUS

Box 7	CATEG
Boxes 8-9	FEM
Boxes 10-11	MAL
Boxes 12-13	NUT
Boxes 14-15	GPFEM
Boxes 16-17	GPMAL
Boxes 18-19	GPNUT
Boxes 20-21	FEMREF
Boxes 22-23	MALREF
Boxes 24-25	NUTREF
Boxes 26-27	GPFEMREF
Boxes 28-29	GPMALREF
Boxes 30-31	GPNUTREF
Box 32	FEMROLF
Box 33	FEMROLM
Box 34	FEMROLN
Box 35	MALROLF
Box 36	MALROLM
Box 37	MALROLN
Box 38	NUTROLF

Box 39	NUTROLM
Box 40	NUTROLN
Box 41	FEMROLFG
Box 42	FEMROLMG
Box 43	FEMROLNG
Box 44	MALROLFG
Box 45	MALROLMG
Box 46	MALROLNG
Box 47	NUTROLFG
Box 48	NUTROLMG
Box 49	NUTROLNG
Boxes 50-51	TOTFEMROL
Boxes 52-53	TOTMALROL
Boxes 54-55	TOTROL
Box 56	STATUS

STEM

Box 57	CATEG
Boxes 58-59	FEM
Boxes 60-61	MAL
Boxes 62-63	NUT
Boxes 64-65	GPFEM
Boxes 66-67	GPMAL
Boxes 68-69	GPNUT
Boxes 70-71	FEMREF
Boxes 72-73	MALREF
Boxes 74-75	NUTREF
Boxes 76-77	GPFEMREF

Boxes 78-79..GPMALREF

Boxes 80-81..GPNUTREF

Box 82...FEMROLF

Box 83...FEMROLM

Box 84...FEMROLN

Box 85...MALROLF

Box 86...MALROLM

Box 87...MALROLN

Box 88...NUTROLF

Box 89...NUTROLM

Box 90...NUTROLN

Box 91...FEMROLFG

Box 92...FEMROLMG

Box 93...FEMROLNG

Box 94...MALROLFG

Box 95...MALROLMG

Box 96...MALROLNG

Box 97...NUTROLFG

Box 98...NUTROLMG

Box 99...NUTROLNG

Boxes 100-101..TOTFEMROL

Boxes 102-103..TOTMALROL

Boxes 104-105..TOTROL

Box 106..STATUS

RESPONSE 1

Box 107..CATEG

Boxes 108-109..FEM

Boxes 110-111	MAL
Boxes 112-113	NUT
Boxes 114-115	GPFEM
Boxes 116-117	GPMAL
Boxes 118-119	GPNUT
Boxes 120-121	FEMREF
Boxes 122-123	MALREF
Boxes 124-125	NUTREF
Boxes 126-127	GPFEMREF
Boxes 128-129	GPMALREF
Boxes 130-131	GPNUTREF
Box 132	FEMROLF
Box 133	FEMROLM
Box 134	FEMROLN
Box 135	MALROLF
Box 136	MALROLM
Box 137	MALROLN
Box 138	NUTROLF
Box 139	NUTROLM
Box 140	NUTROLN
Box 141	FEMROLFG
Box 142	FEMROLMG
Box 143	FEMROLNG
Box 144	MALROLFG
Box 145	MALROLMG
Box 146	MALROLNG
Box 147	NUTROLFG

Box 148	NUTROLMG
Box 149	NUTROLNG
Boxes 150-151	TOTFEMROL
Boxes 152-153	TOTMALROL
Boxes 154-155	TOTROL
Box 156	STATUS
RESPONSE 2	
Box 157	CATEG
Boxes 158-159	FEM
Boxes 160-161	MAL
Boxes 162-163	NUT
Boxes 164-165	GPFEM
Boxes 166-167	GPMAL
Boxes 168-169	GPNUT
Boxes 170-171	FEMREF
Boxes 172-173	MALREF
Boxes 174-175	NUTREF
Boxes 176-177	GPFEMREF
Boxes 178-179	GPMALREF
Boxes 180-181	GPNUTREF
Box 182	FEMROLF
Box 183	FEMROLM
Box 184	FEMROLN
Box 185	MALROLF
Box 186	MALROLM
Box 187	MALROLN

Box 188	NUTROLF
Box 189	NUTROLM
Box 190	NUTROLN
Box 191	FEMROLFG
Box 192	FEMROLMG
Box 193	FEMROLNG
Box 194	MALROLFG
Box 195	MALROLMG
Box 196	MALROLNG
Box 197	NUTROLFG
Box 198	NUTROLMG
Box 199	NUTROLNG
Boxes 200-201	TOTFEMROL
Boxes 202-203	TOTMALROL
Boxes 204-205	TOTROL
Box 206	STATUS

RESPONSE 3

Box 207	CATEG
Boxes 208-209	FEM
Boxes 210-211	MAL
Boxes 212-213	NUT
Boxes 214-215	GPFEM
Boxes 216-217	GPMAL
Boxes 218-219	GPNUT
Boxes 220-221	FEMREF
Boxes 222-223	MALREF

Boxes 224-225	NUTREF
Boxes 226-227	GPFEMREF
Boxes 228-229	GPMALREF
Boxes 230-231	GPNUTREF
Box 232	FEMROLF
Box 233	FEMROLM
Box 234	FEMROLN
Box 235	MALROLF
Box 236	MALROLM
Box 237	MALROLN
Box 238	NUTROLF
Box 239	NUTROLM
Box 240	NUTROLN
Box 241	FEMROLFG
Box 242	FEMROLMG
Box 243	FEMROLNG
Box 244	MALROLFG
Box 245	MALROLMG
Box 246	MALROLNG
Box 247	NUTROLFG
Box 248	NUTROLMG
Box 249	NUTROLNG
Boxes 250-251	TOTFEMROL
Boxes 252-253	TOTMALROL
Boxes 254-255	TOTROL
Box 256	STATUS

RESPONSE 4

Box 257	CATEG
Boxes 258-259	FEM
Boxes 260-261	MAL
Boxes 262-263	NUT
Boxes 264-265	GPFEM
Boxes 266-267	GPMAL
Boxes 268-269	GPNUT
Boxes 270-271	FEMREF
Boxes 272-273	MALREF
Boxes 274-275	NUTREF
Boxes 276-277	GPFEMREF
Boxes 278-279	GPMALREF
Boxes 280-281	GPNUTREF
Box 282	FEMROLF
Box 283	FEMROLM
Box 284	FEMROLN
Box 285	MALROLF
Box 286	MALROLM
Box 287	MALROLN
Box 288	NUTROLF
Box 289	NUTROLM
Box 290	NUTROLN
Box 291	FEMROLFG
Box 292	FEMROLMG
Box 293	FEMROLNG

Box 294	MALROLFG
Box 295	MALROLMG
Box 296	MALROLNG
Box 297	NUTROLFG
Box 298	NUTROLMG
Box 299	NUTROLNG
Boxes 300-301	TOTFEMROL
Boxes 302-303	TOTMALROL
Boxes 304-305	TOTROL
Box 306	STATUS

RESPONSE 5

Box 307	CATEG
Boxes 308-309	FEM
Boxes 310-311	MAL
Boxes 312-313	NUT
Boxes 314-315	GPFEM
Boxes 316-317	GPMAL
Boxes 318-319	GPNUT
Boxes 320-321	FEMREF
Boxes 322-323	MALREF
Boxes 324-325	NUTREF
Boxes 326-327	GPFEMREF
Boxes 328-329	GPMALREF
Boxes 330-331	GPNUTREF
Box 332	FEMROLF
Box 333	FEMROLM

Box 334	FEMROLN
Box 335	MALROLF
Box 336	MALROLM
Box 337	MALROLN
Box 338	NUTROLF
Box 339	NUTROLM
Box 340	NUTROLN
Box 341	FEMROLFG
Box 342	FEMROLMG
Box 343	FEMROLNG
Box 344	MALROLFG
Box 345	MALROLMG
Box 346	MALROLNG
Box 347	NUTROLFG
Box 348	NUTROLMG
Box 349	NUTROLNG
Boxes 350-351	TOTFEMROL
Boxes 352-353	TOTMALROL
Boxes 354-355	TOTROL
Box 356	STATUS